Genome-Scale Modeling and Experimental Approaches to Discover Novel Microbial Metabolic Capabilities

by

Shu Pan

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemical Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2018

Date of final oral examination: 5/29/2018

The dissertation is approved by the following members of the Final Oral Committee:

Jennifer L. Reed, Professor, Chemical and Biological Engineering
Brian F. Pfleger, Professor, Chemical and Biological Engineering
John Yin, Professor, Chemical and Biological Engineering
Ophelia S. Venturelli, Assistant Professor, Biochemistry
Philip A. Romero, Assistant Professor, Biochemistry

To my mentors, colleagues, family, and friends

Acknowledgements

This dissertation would not have been possible without the support and love from my mentors, colleagues, family members, and friends.

First and foremost, I would like to sincerely thank my Ph.D. advisor, Prof. Jennifer Reed for her guidance and mentoring for the past five years. I was inspired by her great dedication to work and perseverance in fighting against difficulties along the way. In addition to Prof. Reed, I would like to thank my mentors and collaborators from the Squid and Vibrio labs, Prof. Edward (Ned) Ruby and Prof. Margaret McFall-Ngai, who introduced me to the fascinating model system of the Squid and Vibrio. Together with Prof. Reed, they encouraged me from the beginning of my Ph.D. career and proved to me that someone could truly be in love with scientific research his/her entire life.

I would also like to acknowledge members of my thesis committee, Prof. John Yin, Prof. Philip Romero, Prof. Ophelia Venturelli, and Prof. Brian Pfleger (also an audience of my group meeting presentations during the semesters), and my preliminary exam and fourth-year talk committee member, Prof. Eric Shusta, for spending their time in learning about my research projects and sharing their insights and expertise.

In the past five years, I have received continuous and generous help from numerous colleagues that I have worked with: Josh Hamilton, Joonhoon Kim, Xiaolin Zhang, Wai Kit Ong, Matt Long, Mingyuan Tian, Paul Adamczyk, Prashant Kumar, Sanjan Gupta, and Hugh Purdy from the Reed group; Kiel Nikolakakis, Min Pan, and Julia Schwartzman from the Squid and Vibrio labs; and members from the Pfleger, Shusta, Palecek, Romero, and Lynn groups at the UW-Madison.

Finally, I have to give my special thanks to my family and friends. The tremendous freedom and support from my parents allowed me to pursue my dreams without worrying about anything else. I was deeply encouraged by my 97-year old grandfather who called me “Dr. Pan” since I was in college, and I felt being loved by my grandmother who treated me as a little girl and never stopped missing me. Last but not least, I am grateful to the joy and happiness brought to me by my dearest friends even when they were far away from me.

Table of Contents

List of Figures

List of Tables

Chapter 1

Current State of Genome-Scale Metabolic Modeling

Some material in this chapter has been adapted from:

> Pan S, Reed JL: **Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries**. *Curr. Opin. Biotechnol.* 2018, **51**:103–108.

A genome-scale metabolic network reconstruction (GENRE) provides a systems-level representation of biochemical reactions within a specific organism and of genes required for these reactions. Due to better and cheaper whole-genome sequencing, the number of reconstructions has rapidly grown since the generation of the first GENRE for *Haemophilus influenza* in 1999 [1]. GENREs and genome-scale metabolic models (GEMs), along with computational algorithms, have led to *in silico* discoveries in many areas, including metabolic engineering, cellular phenotypes, biological network properties, evolutionary processes, interspecies interactions, and systems medicine [2–5]. In **Chapter 1**, basic terminologies used in GENREs and genome-scale metabolic models (GEMs) and algorithms designed for GEMs are first introduced to audience unfamiliar with this research field. Then current applications and advances in gap-filling algorithms, which are key for GENREs and GEMs, are discussed.

1.1 Genome-Scale Metabolic Models and Flux-Calculating Algorithms

1.1.1 Structure of Genome-Scale Metabolic Models

A GEM is the mathematical representation of a GENRE. In a GEM, metabolic reactions are stored in a stoichiometric matrix (S) for each metabolite and reaction, and

genes are connected to reactions by gene-protein-reaction (GPR) associations in the form of Boolean rules (Figure 1.1) [6].

Abbreviation	Glycolytic reactions	Genes
HEX1	[c]GLC + ATP → G6P + ADP + H	*glk*
PGI	[c]G6P ↔ F6P	*pgi*
PFK	[c]ATP + F6P → ADP + FDP + H	*pfkA, pfkB*
FBA	[c]FDP ↔ DHAP + G3P	*fbaA, fbaB*
TPI	[c]DHAP ↔ G3P	*tpiA*
GAPD	[c]G3P + NAD + PI ↔ 13DPG + H + NADH	*gapA, gapC1, gapC2*
PGK	[c]13DPG + ADP ↔ 3PG + ATP	*pgk*
PGM	[c]3PG ↔ 2PG	*gpmA, gpmB*
ENO	[c]2PG ↔ H_2O + PEP	*eno*
PYK	[c]ADP + H + PEP → ATP + PYR	*pykA, pykF*

	HEX1	PGI	PFK	FBA	TPI	GAPD	PGK	PGM	ENO	PYK
ATP	−1	0	−1	0	0	0	1	0	0	1
GLC	−1	0	0	0	0	0	0	0	0	0
ADP	1	0	1	0	0	0	−1	0	0	−1
G6P	1	−1	0	0	0	0	0	0	0	0
H	1	0	1	0	0	1	0	0	0	−1
F6P	0	1	−1	0	0	0	0	0	0	0
FDP	0	0	1	−1	0	0	0	0	0	0
DHAP	0	0	0	1	−1	0	0	0	0	0
G3P	0	0	0	1	1	−1	0	0	0	0
NAD	0	0	0	0	0	−1	0	0	0	0
PI	0	0	0	0	0	−1	0	0	0	0
13DPG	0	0	0	0	0	1	−1	0	0	0
NADH	0	0	0	0	0	1	0	0	0	0
3PG	0	0	0	0	0	0	1	−1	0	0
2PG	0	0	0	0	0	0	0	1	−1	0
PEP	0	0	0	0	0	0	0	0	1	−1
H_2O	0	0	0	0	0	0	0	0	1	0
PYR	0	0	0	0	0	0	0	0	0	1

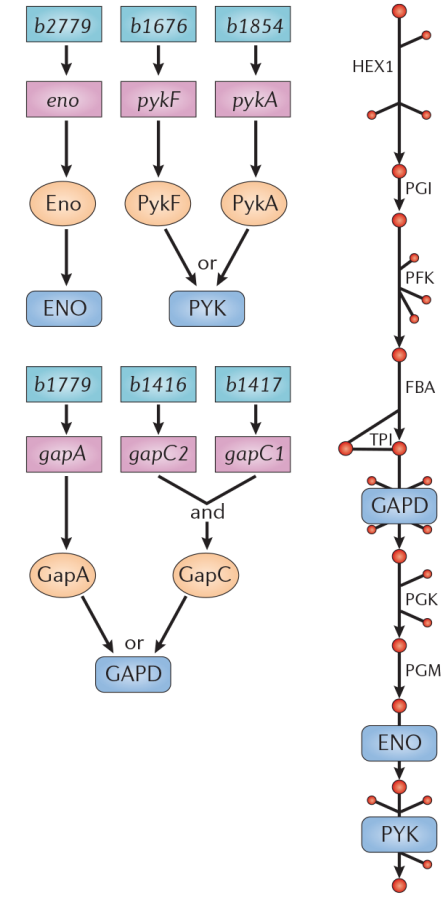


Figure 1.1 Example of a Stoichiometric Matrix (*S*) and GPR for a Short Metabolic Pathway. All reactions in this short pathway are explicitly written out in charge and elementally balanced format and stored in *S*. Enzymes catalyzing each reaction and Boolean rules for GPR associations are also identified. For example, GapA or GapC could catalyze reaction GAPD. GapA is encoded by a single gene, *gapA*, while GapC is encoded by both *gapC1* and *gapC2* [6].

1.1.2 Flux Balance Analysis and Forced Coupling Algorithms

Flux balance analysis (FBA) is a linear optimization algorithm that calculates reaction fluxes in a GEM that maximize an objective, e.g. cell growth, under the steady state assumption of a metabolic network. At steady state, there is no accumulation of intermediate metabolites, and the rest of the products contribute to biomass. Therefore,

FBA solves a system of linear equations with reaction fluxes (v) as variables, i.e., $S \cdot v = 0$. In general, there are multiple solutions to this system of equations because $S \cdot v = 0$ is an underdetermined system. In addition to stoichiometric constraints, enzyme-capacity $(\alpha \leq v_j \leq \beta)$ and thermodynamics constraints $(0 \leq v_j)$ could reduce the number of solutions subject to one or multiple objectives (See Figure 1.2).

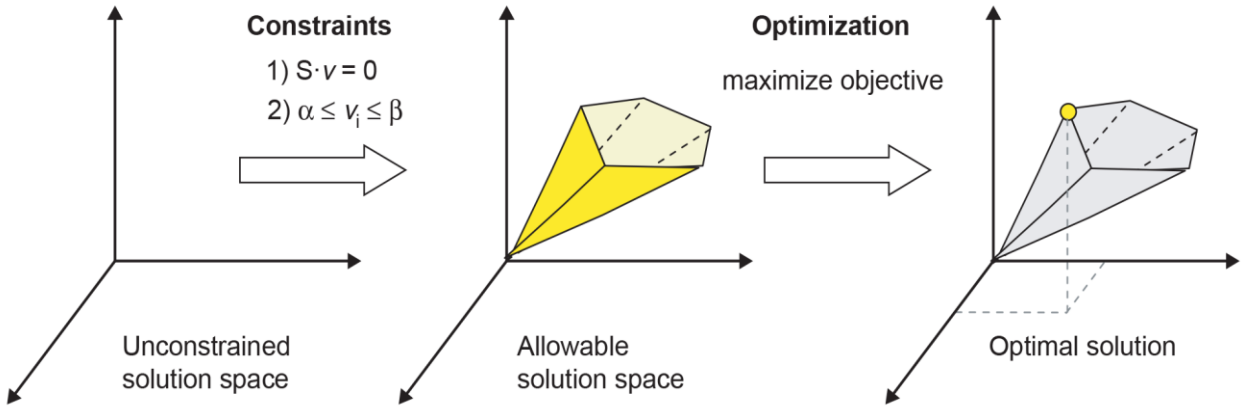


Figure 1.2 Schematic of How FBA Works. A three-dimensional solution space represents all possible values that variables can take on when the problem is unconstrained. When stoichiometric, enzyme-capacity, and thermodynamic constraints are applied, the solution space is reduced to the allowable solution space, a convex cone. Then, a linear optimization solver could identify a solution in the cone for a specific objective function. (Figure adapted from [7].)

Forced coupling algorithms—through gene and reaction deletions or additions, and gene over or under-expression—propose design strategies of strains that have coupled reaction fluxes [8–12]. For example, OptORF[8], a bi-level programming algorithm, couples the fluxes of cell growth and chemical production, thus creating a positive correlation between these two fluxes by deleting or overexpressing metabolic or regulatory genes. Therefore, the desired chemical production becomes faster when the strain is operating towards its maximal growth rate. Experimentally, the maximal growth is likely to be achieved by adaptive laboratory evolution (ALE), where cells are passaged

at mid-exponential phase for multiple generations to acquire beneficial mutations. More recently, the ShadowCon algorithm by Tervo and Reed has further extended previous forced coupling algorithms [8,10,13,14] to finely control how tightly two reaction fluxes are correlated [15]. Figure 1.3 shows that when added to OptORF, ShadowCon ensures that the solution has a degree of coupling between user-specified lower (m^{min}) and upper (m^{max}) bound.

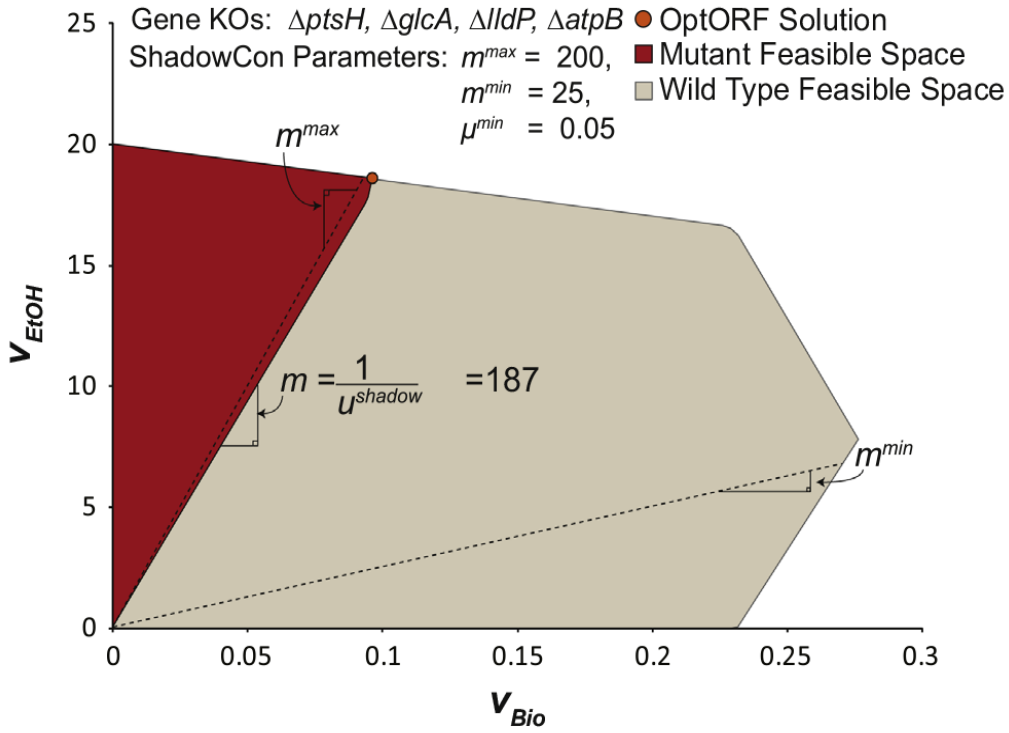


Figure 1.3 Example of a ShadowCon Solution. The ShadowCon limits the slope of the solution space when coupling between ethanol and biomass production occurs. The dual variable, u^{shadow}, constrains the slope, m, between a user-specified upper (m^{max}) and lower (m^{min}) bound (dashed lines) [15].

1.2 Advances in Gap-Filling Genome-Scale Metabolic Models

A draft GEM is often generated automatically by software platforms, which use genome annotations of a specific organism and connect genes to metabolic reactions using reference databases [16,17]. A draft model has to be further refined and evaluated in

multiple steps to ensure its quality [16,17]. This refinement and evaluation process includes gap-filling, which improves the network connectivity by modifying content of the metabolic model. Gap-filling analyses can lead to discoveries of missing reactions, unknown pathways, unannotated and misannotated genes, as well as promiscuous enzymes and underground metabolic pathways. Classic gap-filling algorithms have been reviewed previously by Orth and Palsson [18]. These algorithms generally include three steps: detecting gaps, changing model content to fill gaps (i.e., add/remove reactions, change biomass compositions, or change reaction reversibility), and identifying genes responsible for the gap-filled reactions (Figure 1.4).

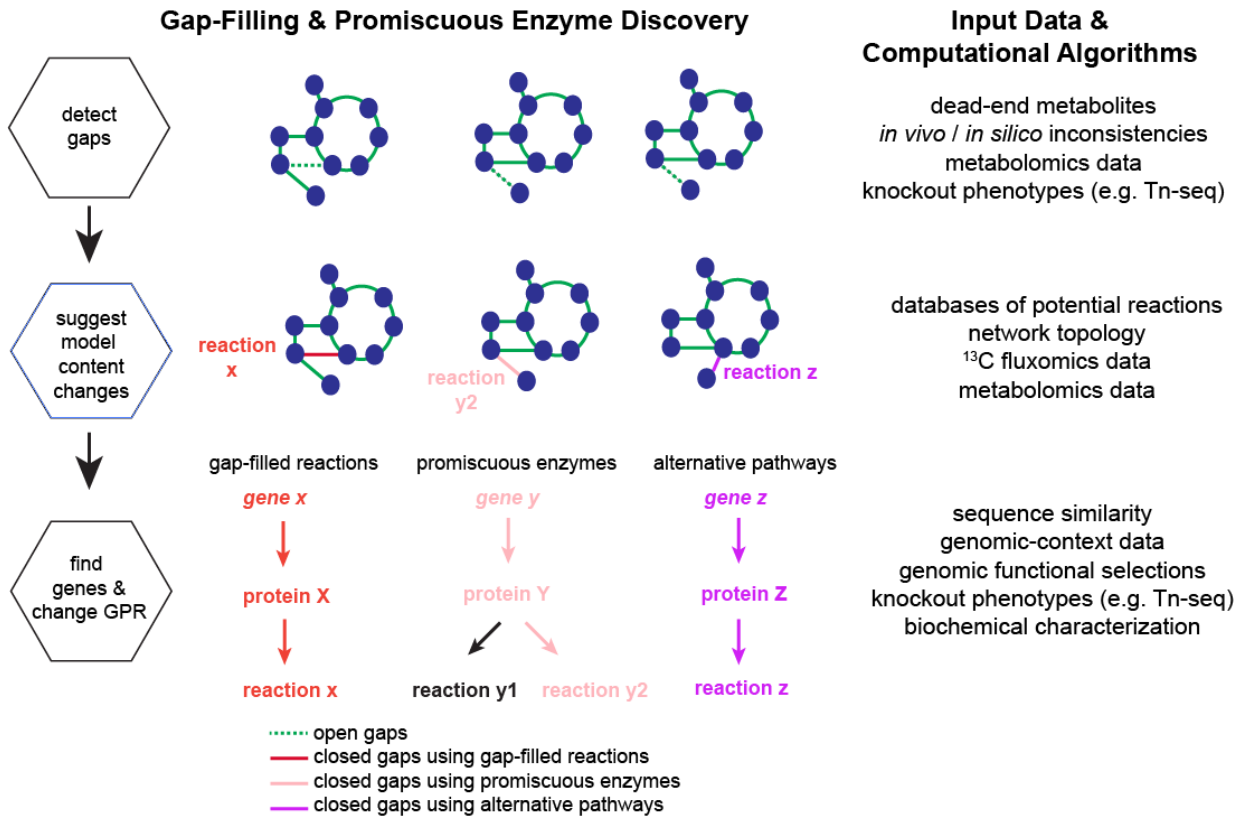


Figure 1.4 Steps and Input Data of Gap-Filling. First, dead-end metabolites and *in silico* and *in vivo* inconsistencies allow detection of gaps in metabolic models. Then, the model content (i.e., reactions and biomass compositions) is changed to resolve these inconsistencies. In this step, missing reactions can be added from databases, and

network topology analysis can rank these potential reactions. Finally, the genes responsible for the filled gaps are identified using sequence similarity, genomic-context data, and genomic functional selections and are verified by biochemical characterizations. Similarly, promiscuous enzymes and underground metabolic pathways can also be identified when analyzing the gaps in the models.

In the first step, gap-filling algorithms identify dead-end metabolites (metabolites which cannot be consumed or produced in the network), and/or inconsistencies between model predictions and experimental data (e.g., growth phenotypes). They then solve for a set of reactions from metabolic databases of potential reactions that if added to the metabolic model "activate" dead-end metabolites or resolve the inconsistencies. In the third step, some gap-filling algorithms discover genes that could be responsible for these reactions, which can be further tested biochemically or genetically. A simple gap-filling example is illustrated in Figure 1.5.

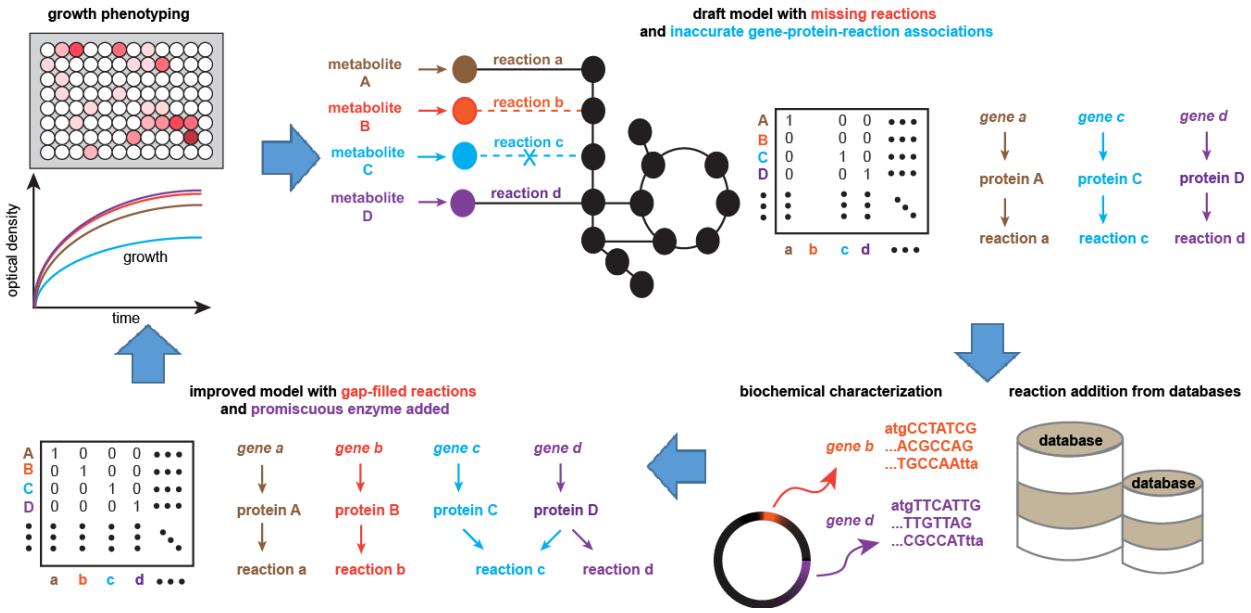


Figure 1.5 An Example of the Iterative Process of Laboratory and Computational Experimentation in Gap-Filling. As revealed by high-throughput growth phenotyping experiments, the wild type can grow on metabolites A, B, or D as a sole carbon source. Additionally, a knockout mutant of *gene c* that known to catalyze reaction c can grow on metabolite C. Contrarily, the model incorrectly predicts that the wild type cannot grow on B and the *gene c* mutant cannot grow on C, indicating that reaction b is missing from the model and there exists another enzyme that can catalyze reaction c. To fix these inconsistencies, gap-filling adds reaction b from databases and finds that *gene d* has

sequence similarity to *gene c*. Further biochemical characterizations confirm that protein b is responsible for reaction b and protein d has weak activities towards reaction c. The model is then updated and compared to additional experimental datasets to allow iterative metabolic discoveries and to improve prediction accuracy.

1.2.1 Advances in Gap-Detection and Reaction-Addition Algorithms

Some recent algorithms aim to detect and fill gaps more efficiently than earlier gap-filling algorithms [18]. For example, FASTGAPFILL [19] is a scalable algorithm that computes a near minimal set of added reactions for a compartmentalized model. Another method, GLOBALFIT [20], reformulates the mixed integer linear programming problem of gap-filling into a simpler bi-level linear optimization problem. It efficiently identifies the minimal set of network changes needed to correct multiple *in silico* predictions that are inconsistent with *in vivo* observations simultaneously. Meneco [21] and a hybrid metabolic network completion algorithm [22] reformulate the reaction-addition problem using answer set programming, a declarative programming paradigm intended to solve difficult combinatorial search problems. Their usage of answer set programming allows for stoichiometry constraints to be violated, potentially resulting in solutions which are less biased by the inaccurate stoichiometry of a model. The hybrid metabolic network completion approach combines answer set programming with linear stoichiometry constraints, and offers a better solution for restoring highly degraded models [22].

Algorithms, such as GAUGE [12], have been developed to exploit alternative mechanisms for gap-identification. GAUGE exploits flux coupling analysis (FCA) [13] that detects how two reactions depend on each other. Using FCA, GAUGE finds gaps involving genes that are associated with fully dependent reactions but show uncorrelated expression patterns. However, GAUGE can only analyze a subset of a model where

gene-protein-reaction (GPR) associations are defined and isozymes or multi-functional genes do not create possibilities for uncorrelated gene expression patterns.

Some algorithms exploit alternative mechanisms for adding reactions. The novel algorithm DEF [24] is based upon filling reactions to “activate” dead-end metabolites in a manner similar to eukaryotes engulfing mitochondria to find the most efficient pathways for consuming oxygen. Following this quasi-endosymbiosis theory, DEF aims to add reactions that maximize production/consumption of dead-end metabolites in the original model. A DEF solution could contain more reactions that are biologically reasonable compared to a parsimonious gap-filling solution, which often is a minimal set of reactions.

Inherently different from all algorithms mentioned above, pattern-based gap-filling algorithms do not contain an explicit gap-identification or reaction-addition step. In a metabolite pattern and probabilistic method [25], feature propagation Markov models (HMMs) are used to rank potential gap-filled reactions by how closely they are related to the network. In MATBoost and BoostGAPFILL [16, 17], a training incidence matrix, S, with artificial gaps is created by deleting reactions randomly from a network. Then a machine learning technique, matrix factorization, completes the missing entries creating matrix A. Finally, an integer least square optimization selects reactions from a database that best match A. These steps, which are based on the network pattern alone, cannot fill in all the gaps in the network. However, they can rank and set weights of reactions in databases and work in conjunction with other gap-filling algorithms such as FASTGAPFILL [19] to improve their gap-filling accuracy.

Unlike the above algorithms that add reactions to an incomplete metabolic model of any organism, SONEC [18] works for an organism in a metagenomics sample. It is

initialized with a bin of contigs for each organism in the sample and a bin for unassigned sequence fragments. Then it computes metabolite connectivity scores between the unassigned fragments and each organism to determine which organism these fragments belong to.

The abovementioned gap-detection and reaction-addition algorithms are used in the first two steps of gap-filling (Figure 1.4). They can hypothesize new network components including gap-filled reactions. After the two steps, genes are assigned to gap-filled reactions (Figure 1.4). This last step of gap-filling is a very important yet under-explored step which should be supported with additional experiments.

1.2.2 Advances in Gene-Assignment Algorithms and Approaches

Knowing the gene sequence(s) responsible for a reaction is extremely useful in applications using genetic manipulation or drug discoveries. Several algorithms that have been previously reviewed [18] provide bioinformatics solutions to match gene sequences to gap-filled reactions. These methods could incorporate data such as sequence similarity, co-expression, chromosomal proximity, and phylogenetic profiles. Recently, GLOBUS [29] has combined these data with a global probabilistic approach to provide probable annotations including possible alternate functions. Another novel algorithm GO-MEP [30] identifies biologically relevant groups of proteins (e.g. proteins in the same pathways, protein complexes, proteins with the same localization, and physically interacting protein pairs) and has showed improved accuracy of identifying missing genes.

An alternative approach for gene assignment is to directly incorporate gene assignment in the gap-filling procedure. In MIRAGE [31], phylogenetic profiles and gene expression are first used to estimate the likelihood that a specific reaction in the database

can be added. In the next step, starting from a model with all reactions in the database added, a new model is created to determine the set of gap-filled reactions by removing low likelihood reactions while keeping the desired model properties. Another likelihood-based gap-filling algorithm [32] and its web service (ProbAnnoWeb) and standalone python package (ProbAnnoPy) [33], have been developed to directly incorporate gene assignment into gap-filling. In ProbAnnoWeb and ProbAnnoPy, an organism-specific likelihood score is assigned to each reaction in a database of potential reactions. These scores are calculated based on the BLASTp results of a gene in the organism genome against high confidence functional annotations. Reactions with higher scores are weighted heavier in the objective function. Therefore, a ProbAnno solution might contain more reactions than a parsimonious gap-filling solution. Similar to ProbAnno, there exists another BLAST-weighted gap-filling algorithm [34] in which reactions are added from weighted biochemistry databases.

Ultimately, all gap-filled gene-reaction associations should be evaluated in subsequent laboratory experiments. For example, a change in gene expression might be measured by quantitative PCR (qPCR) or growth phenotypes of knockout mutants under different media conditions. Or the product of an enzyme's proposed reaction could be measured by analytical chemistry methods.

1.2.3 Combining Gap-Filling Algorithms with High-Throughput Phenotyping Experiments

Laboratory experiments are not only important to validate gap-filling solutions, but also are important for identifying the metabolic gaps themselves. Cheaper and more available high-throughput experimental datasets, which can be compared to *in silico* predictions, have also made gap-filling analyses increasingly powerful [35,36]. For

example, gap-filling algorithms could improve the quality of a metabolic model by resolving incorrect growth phenotype predictions for knockout mutants under different medium conditions. Currently, phenotypic microarrays and robotic instruments [37] can test growth of cells in fifty 96-well plates for hundreds of different media conditions. RB-TnSeq has also allowed rapid quantification of genome-wide mutant fitness in 387 successful assays [21]. With improved genome editing tools such as MAGE [39,40], pORTMAGE [41], and CRMAGE [42], more knockout libraries will be added to the current collections [43–49] in the near future. Such libraries can be rapidly tested under a variety of conditions to identify inconsistencies between model predictions and experimental observations and to identify metabolic gaps.

Chapter 2

A Genome-Scale Metabolic Reconstruction *i*VF846 That sheds Light on the Host-Associated Metabolism of *Vibrio fischeri*

Vibrio fischeri is a bioluminescent Gram-negative marine bacterium that can form a mutualistic relationship with some squids or fishes [50]. In the 1960s the bioluminescent property of *V. fischeri* was discovered [51] and later quorum sensing was revealed to control bioluminescence in a cell density dependent manner. Since then, there has been a long-standing interest in *V. fischeri* for studying bioluminescence, quorum sensing, and host-microbe interactions. For example, researchers have found that the interactions between *V. fischeri* and Hawaiian bobtail squid *Euprymna scolopes* follow daily rhythmic cycles. At dawn the squid expels 90% of the *V. fischeri* and the remaining cells grow on nutrients fed by the squid. By night, *V. fischeri* cells reach a peak density that enables them to produce bioluminescence for the squid. The *V. fischeri* ES114 strain, a symbiotic isolate from the squid [52], has become the experimental strain for many research studies. A plethora of tools for molecular and systems biology [53–55], labeling, and spectroscopy [56] have been applied or developed to study ES114 and other *V. fischeri* strains and their interactions with their hosts.

The *V. fischeri* metabolism plays a key role in squid-Vibrio interactions. The diverse metabolic capabilities of *V. fischeri* enable it to adapt to drastically different environmental conditions, e.g., living inside or outside its squid host, and when it is a dimly lit individual cell or part of a bright dense bioluminescent group. Researchers have found mechanisms that involve specific nutrient sources, chemical cues, and regulatory proteins that actively

control the metabolic pathways that *V. fischeri* utilizes when free-living and host-associated [50]. Such knowledge is transferable to understand more complex host-microbe interactions, e.g., human microbiome interactions, and key for developing new antibiotics. However, details of many mechanisms that contribute to the diverse and flexible metabolic capabilities of *V. fischeri* are still unknown. One promising strategy that could facilitate the investigation of these mechanisms is the metabolic network reconstruction, which has been demonstrated to be useful in areas such as model-driven discovery, prediction of cellular phenotype, and analysis of biological network properties [4,5].

Here we report the genome-scale metabolic network reconstruction (GENRE) for *V. fischeri* ES114, *i*VF846. To our knowledge, it is the first manually curated *V. fischeri* GENRE and the second manually curated Vibrio GENRE [57]. This reconstruction has been used for suggesting nutrient sources provided to ES114 by its squid host [53] and discoveries of unknown genes associated with knowledge gaps in the *V. fischeri* metabolic network [58]. Through the comparison between the GENREs of *V. fischeri* and *Escherichia coli*, i.e., *i*VF846 and *i*JO1366 respectively, and the comparison between *i*VF846 and the genomic information of 12 other *V. fischeri* strains of either “dominant” or “sharing” type of colonization phenotypes [55], we uncovered some general patterns and details about the unique metabolic capabilities of *V. fischeri* that are potentially important for Vibrio-squid interactions. Finally, we demonstrated bioluminescence could be a significant energy cost to *V. fischeri* especially under limited oxygen conditions using *in silico* simulation with *i*VF846.

2.1 Results and Discussion

2.1.1 Reconstructing a Metabolic Network for *V. fischeri* ES114

After obtaining a draft model (see Appendix I Methods for details), high-throughput experimental datasets [54,58] that include growth phenotypes of wildtype *V. fischeri* ES114 on sole carbon sources and gene essentiality data were used to further improve and validate the draft model. We used a gap-filling algorithm to add reactions from *i*JO1366 so the experimental growth phenotypes matched the computational predictions as much as possible [59]. After searching for possible *V. fischeri* genes responsible for the gap-filled reactions in literature, NCBI [60], MetaCyc [61], and KEGG [62] databases, we did find not good hits for the following reactions in Table 2.1. Six of the reactions were transport reactions and the rest were involved in cofactor and prosthetic group biosynthesis, cell envelope biosynthesis, alternative carbon metabolism, and folate metabolism. However, we performed experiments and found three *V. fischeri* genes that have no sequence similarity to *E. coli* genes to be responsible for some of these reactions [58]. After finding the three *V. fischeri* genes, there exist only seven non-GPR reactions gap-filled into the *V. fischeri* model and majority them (five out of seven) are transport reactions, which usually exist in GENREs due to poor annotation of transporters. This shows that we generally have high confidence for *i*VF846 reactions from the support of genomic evidence.

Reaction Name	*E. coli* Genes	Subsystems	Experimentally Discovered *V. fischeri* Genes
aspartate 1-decarboxylase proenzyme	b0131	Cofactor and Prosthetic Group Biosynthesis	VF_0892
nucleoside-specific channel-forming protein	b0411	Transporter	
putative pyrimidine:cation symporter	b1006	Transporter	
enoyl-[acyl-carrier-protein] reductase	b1288	Cell Envelope Biosynthesis	
aldehyde dehydrogenase A, NAD-linked	b1415	Folate Metabolism	
xanthosine:H^+ symporter	b2406	Transporter	
N-acetylneuraminate:H^+ symporter	b3224	Transporter	VF_0668
mannitol-1-phosphate 5-dehydrogenase	b3600	Alternate carbon metabolism	VF_A0062
(R)-lactate/(S)-lactate/glycolate:H^+ symporter	b3603	Transporter	
maltose outer membrane porin / phage lambda receptor protein	b4036	Transporter	

Table 2.1 Non-GPR Reactions Gap-Filled to the Draft Model. Six are transport reactions and three have *V. fischeri* genes later discovered to be responsible for these reactions [58].

A manual curation process was used for resolving some of the mismatches between model-predicted and experimentally-suggested essential genes. We were able to add exchange reactions, modify GPR rules, and find isozymes or alternative pathways to resolve some of the mismatches. We also designed experiments and discovered glutamine transporter genes (VF_0924 and VF_1172) that originally was missing from the draft model due to poor annotation [58].

Several updates were made upon the initial model that we used for the transcriptomics analysis and gene discovery projects [53,58]. Previously, all exchange reactions from *i*JO1366 were kept in *i*VF846. In this updated version, only the exchange

reactions that are connected to at least another metabolic reaction were kept. Additionally, three non-GPR reactions previously copied from *i*JO1366, i.e., 2-methylcitrate dehydratase, methylisocitrate lyase, and 2-methylcitrate synthase encoded by *prpBCDE* in the PrpR regulon, were deleted due to missing genomic evidence. These methyl citrate cycle reactions are involved with propionyl-CoA metabolism. Therefore, *E. coli* can process propionyl-CoA from degradation of odd-chain-length fatty acids, and it can also grow on propionate as a sole carbon source although slowly [63]. In contrast, *V. fischeri* showed no growth on propionate as a sole carbon source. In addition to the no-growth phenotype, *V. fischeri* ES114 and at least two other Vibrionales strains that have no PrpR regulon are all missing *mqo*, a gene that could covert the malate from the methyl citrate cycle into oxaloacetate [64]. Although genomic evidence hints at a missing methyl citric cycle, *V. fischeri* ES114 attracts to short- and medium-chain-length fatty acids with backbone lengths of C3 to C8, including the old-chain-length propionate, valerate, and heptanoate [65]. It is interesting why *V. fischeri* attracts to these fatty acids if they are not used as a source of carbon.

Finally, we ensured that the final model does not contain erroneous energy-generating cycles that synthesize energy metabolites such as ATP, and validated our final model using the growth phenotypes and gene essentiality datasets. We compared *in silico* and *in vivo* growth phenotypes of *V. fischeri* ES114 on sole carbon sources under both aerobic and anaerobic growth conditions. Out of the 49 experimentally growth cases, 48 are correctly predicted by the model, and out of the 91 experimentally no-growth cases, 75 are correctly predicted. The accuracy of overall model predicted growth phenotypes is 87.9%, with almost 100% accuracy in predicting growth conditions (Figure 2.1). The

only growth case predicted to be no-growth is the aerobic growth on glycogen. We suspect that glycogen could be degraded into monomers or dimers by periplasmic enzymes before being transported inside the cells. The model is only 82.4% accurate in predicting no-growth cases (Figure 2.1). One of the reasons for these discrepancies might be poor transport of some carbon sources or small carbon fluxes available towards biomass production, which cannot support a fast enough growth to be observed experimentally. There might also be medium components that were not at an optimal concentration in the high-throughput growth phenotyping experiments, and therefore creating more experimentally negative results than the model predictions. Such an example is the growth of *V. fischeri* on glycerol. We observed anaerobic growth of *V. fischeri* in culture tubes on glycerol if we added appropriate levels of $FeSO_4$, nitrate, and fumarate (Table 2.2). There might also exist regulatory restrictions that prevented some chemicals to serve as a sole carbon source for *V. fischeri*.

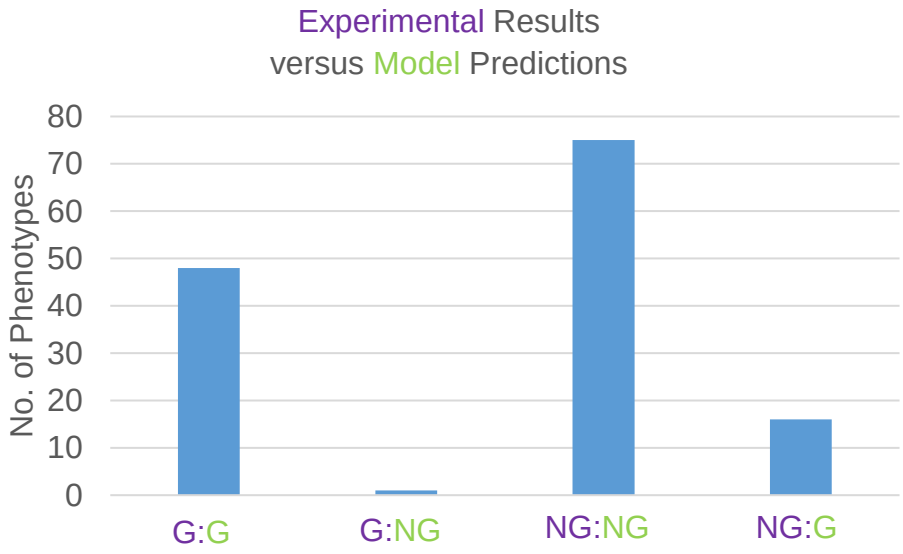


Figure 2.1 Experimental Results and Model Predictions of Wild Type Growth on Sole Carbon Sources. (G: growth, NG: no-growth; purple: experimental results, green: model predictions).

Growth	Carbon source	Electron acceptor	FeSO4
YES	40 mM Glycerol	40 mM TMAO	5 µM FeSO4
YES	40 mM Glycerol	40 mM Fumarate	5 µM FeSO4
NO	40 mM Glycerol	0.4 to 40 mM Nitrate	None
YES	40 mM Glycerol	0.4 to 40 mM Nitrate	5 µM FeSO4
NO	40 mM Glycerol	None	None

Table 2.2 Anaerobic Growth Conditions Associated with *V. fischeri* in DMM Containing Glycerol as a Sole Carbon Source. Appropriate concentrations of electron acceptors and $FeSO_4$ are required for cell growth.

Accuracy in gene essentiality predictions is another benchmark for measuring the quality of a genome-scale metabolic network. Note that some of the experimentally essential *V. fischeri* genes are not essential in *E. coli* or vice versa due to their differences in GPR associations or components of biomass. Therefore GPR associations and biomass composition that originally copied from *i*JO1366 were modified to increase the model's accuracy in gene essentiality predictions. The overall accuracy of gene essentiality predictions by our final model is 91.6% (Figure 2.2). Experimentally, 81.0 % of the 846 genes are nonessential, and the model is very accurate in predicting these nonessential genes with a 96% accuracy.

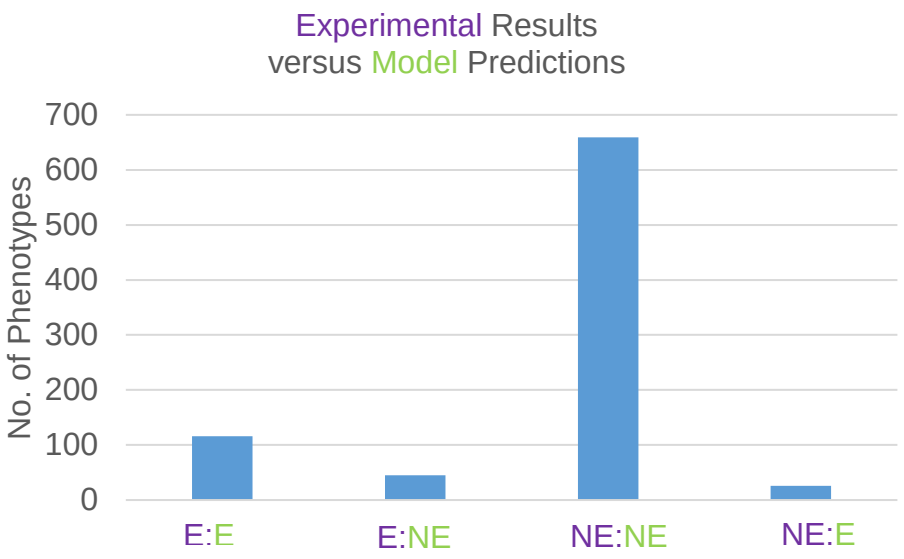


Figure 2.2 Experimental Results and Model Predictions of Gene Essentiality in LBS Medium. (E: essential, NE: none-essential; purple: experimental results, green: model predictions).

2.1.2 Comparing Genome-Scale Metabolic Networks of *E. coli* and *V. fischeri*

*E. coli i*JO1366 was used as the starting point of the *i*VF846 because *V. fischeri* ES114 and *E. coli* MG1665 share many common metabolic pathways. This is supported by the fact that only seven gap-filled reactions were added to the *V. fischeri* draft model from *i*JO1366 to resolve the discrepancies of *in silico* and *in vivo* growth phenotypes.

Knowing the two organisms are similar in many ways, we were interested in the differences *V. fischeri* and *E. coli* metabolism represented by the two models. We calculated the number of unique *V. fischeri* and *E. coli* reactions involved in each category of a general metabolic function. (We considered only GPR reactions in our analysis because non-GPR reactions such as exchange, sink, spontaneous or gap-filled reaction are not supported by genomic evidence.) Out of the 659 unique *E. coli* GPR reactions, about a third (221 reactions) are transport reactions, about 20% (132 reactions) belong

to the alternate carbon metabolism, and another 18% (120 reactions) are related to membranes/lipids/lipopolysaccharides (LPS) biosynthesis (Figure 2.3). Out of the 54 unique *V. fischeri* reactions (encoded by 84 genes), about 22% (12 reactions) are for alternate carbon metabolism, another 22% for amino acid metabolism, 19% (10 reactions) for membranes/lipids/LPS biosynthesis, and 11% (6 reactions) for transport (Figure 2.4).

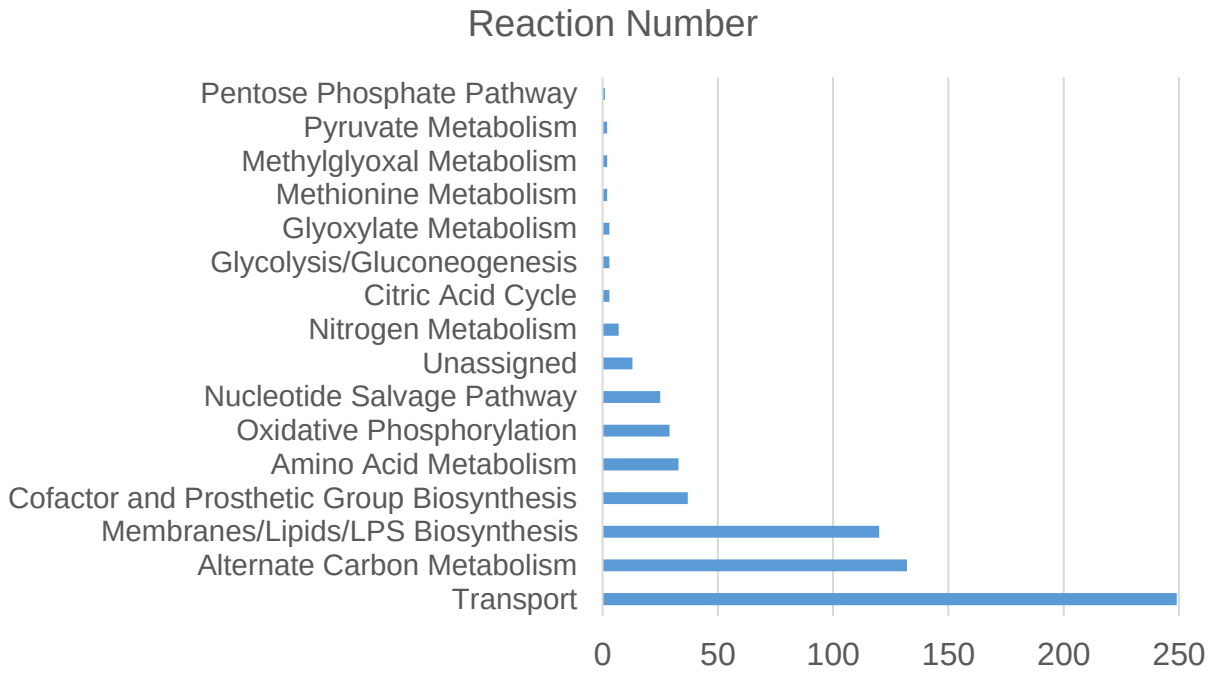


Figure 2.3 GPR Reactions in iJO1366 That Do Not Exist in *i*VF846 Grouped by Subsystems.

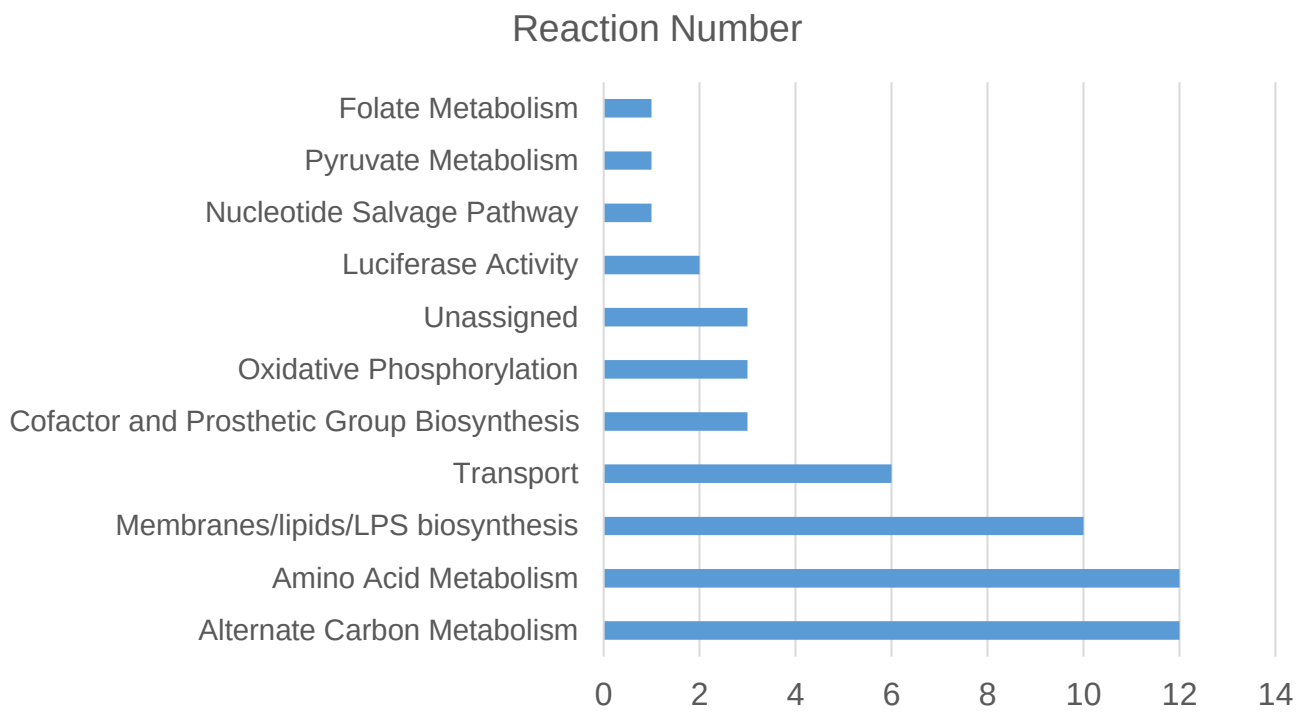


Figure 2.4 GPR Reactions in *i*VF846 That Do Not Exist in *i*JO1366 Grouped by Subsystems.

While many central metabolism reactions, including similar glycolysis, pentose phosphate pathway, and citric cycle reactions, exist in both organisms, a great percentage of unique reactions are for alternate carbon metabolism (20% of all unique reactions in *E. coli* and 22% of all unique reactions in *V. fischeri*) or amino acids metabolism (15% of all unique reactions in *E. coli* and 22% of all unique reactions in *V. fischeri*). This indicates that the two organisms encounter diverse nutrient conditions in their natural habitats where they interact with their hosts. In terms of alternate carbon metabolism, *V. fischeri* does not catabolize the *E. coli*'s sole carbon sources including D-trehalose, L-arabinose, D-xylose, D-lactose, D-allose, and melibiose, L-rhamnose. Furthermore, *V. fischeri* does not contain or only contains incomplete *E. coli* pathways for utilizing substrates like fucose, ketogluconate, glyoxylate, glycolate, D-galactarate, D-glucarate, phenylethylamine, phenylacetate, ornithine, putrescine, 3-phenylpropanoate,

3-(3-hydroxyphenyl)propanoate, cinnamate, and 3-hydroxycinnamate. *V. fischeri* contains many unique reactions in catabolism pathways for chitin sugars, D-galactosamine, and *N*-acetyl-D-galactosamine, and 5-dehydro-4-deoxy-D-glucuronate. Many of these sugars or complex polymeric carbohydrates either have shown great influence on the establishment of *V. fischeri*-squid association or have been found in squid mucins [66–68]. In terms of amino acid metabolism, *V. fischeri* has distinct ways to degrade histidine, glutamate, alanine, arginine, and aspartate. Its unique histidine degradation pathway has an end product of glutamate. In our transcriptomics experiment, this pathway was down-regulated in wild type when *N*-acetyl-D-glucosamine was added (**Chapter 3.1.2**).

Another large fraction of organism-specific reactions are transport reactions (33% of all unique reactions in *E. coli* and 11% of all unique reactions in *V. fischeri*), which are necessary to take up each organism's unique carbon and other nutrient sources. Additionally, a large number of these reactions involved with membranes/lipids/LPS biosynthesis, especially LPS biosynthesis, exist in *E. coli* and *V. fischeri*, and these reactions probably arise from the interactions between the two organisms and their hosts.

Although oxidative phosphorylation pathways do not contain a large number of unique reactions in *V. fischeri* or *E. coli*, we observe some interesting differences in the two organisms. In *E. coli*, the proton pumping NADH dehydrogenase-I (Ndh-I are encoded by *nuo* genes) is required for the anaerobic respiration of NADH using fumarate or DMSO while NADH dehydrogenase II (NDH-II) that does not pump protons is used preferentially in aerobic and nitrate respiration [69]. *V. fischeri* has Ndh-II but not Ndh-I, it also has a unique Na^+ translocating NADH dehydrogenase that is likely used in the salty seawater.

When comparing the cytochromes used by the two organisms, we found *V. fischeri* lacks cytochrome *bo* oxidase (Cyo ABCD), an oxidase that is expressed in *E. coli* when the oxygen level is high. In contrast, the cytochrome *bd* oxidase (CydAB, also present in *E. coli*) and the unique cbb3 cytochrome oxidase (CcoNOQP) of *V. fischeri* both have high affinities for oxygen [50], which enable them to function under oxygen limiting conditions and compete with the luciferase used for bioluminescence. In addition, *V. fischeri* has a predicted alternative oxidase (AOX) which might be associated with response to nitric oxide [50]. Although missing NDH-I and Cyo ABCD, *V. fischeri* has proline/succinate/glycerol 3-phosphate dehydrogenases, Ndh-II, and pyruvate oxidase, which could work in concert with its cytochromes for a variety of electron acceptors.

2.1.3 Comparing Metabolic Capabilities of ES114 and 12 Other *V. fischeri* Strains

To understand the metabolic difference between ES114 and other *V. fischeri* strains, we referred to a previous genomic comparison study of 13 symbiotic *Vibrio fischeri* isolates [55]. According to this study, the genomic structures of these *V. fischeri* strains are highly conserved over time and geographical distance and the core genome common to all strains represent an average of 80.4% of the total proteins encoded [55]. When comparing the genes in *i*VF846 to the core genome from this study, we found about 95% of genes in the model are also in the core genome while only 5% (44 genes) are unique to ES114.

Furthermore, 17 of the 44 unique ES114 genes encode proteins that have isozyme(s) that are in the core genome. Therefore although these 17 genes are not in the core genome, they are responsible for reactions that exist in all *V. fischeri* strains.

Regulation might play a role in controlling in the different isozymes for these 17 reactions under different environmental conditions.

The rest 27 genes are responsible for reactions that do not exist in the core genome involve in the metabolic functions of substrate transport (seven), LPS biosynthesis (four), and electron transport (four). One particularly interesting gene that is unique to ES114 is VF_1301 that encodes a cbb3 cytochrome oxidase component CcoQ because all three other Cbb3-type cytochrome oxidase subunits (VF_1299, VF_1300, and VF_1302, encode CcoNOP) belong to the core genome. Although the functional role of ES114 CcoQ has not been studied, CcoQ is usually found in nitrogen fixing bacteria, also known as FixQ, where it stabilizes the interaction between CcoP and CcoNO against high oxygen concentration [70,71]. Because the ES114 *ccoNOQP* operon has been shown to be down-regulated by chitin sugar catabolism likely for directing oxygen to the luciferase instead of the oxidase [72], we suspect the ES114 CcoQ might help stabilize CcoNOQP complex when chitin sugars and luciferase activity are absent and more oxygen is available for the use by CcoNOQP (Figure 2.5).

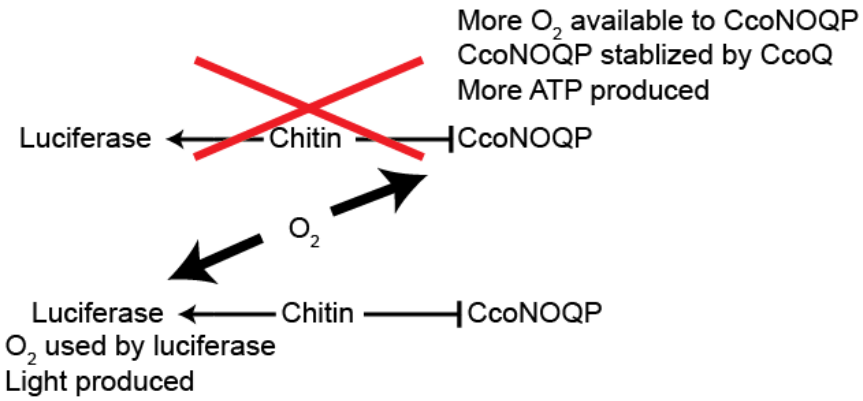


Figure 2.5 Schematic Showing the Hypothesized Role of CcoQ in *V. fischeri*

When mapping these 27 genes, which are responsible for unique ES114 metabolic functions, to the experimentally suggested 73 candidates of colonization factors in *i*VF846

[54], we found that most of the experimentally suggested colonization factors belong to the core genome. The only two out of the 27 genes that might be colonization genes are two transporter genes (VF_2366 for inorganic ion transport and VF_A0529 vitamin B3 transport). Therefore, a large number of squid colonization genes in *V. fischeri* with metabolic functions are conserved in all 13 strains. This is consistent with the previous conclusion that the core genomic structure of *V. fischeri* strains is highly conserved over time and geographical distance.

Despite the similarity in the genomic structure of ES114 and 12 other *V. fischeri* strains, ES114 is an S-type strain that exhibits a strong “sharing” property during squid colonization where two “sharing” strains co-colonize the squid. On the opposite, “dominant” strains (D-type strain) cannot co-exist with another strain in pairwise competition assays. In terms of the genomic structure, the genomes of D-type strains are in general bigger than those of S-type strains. About ~250 Kb of sequence (194 proteins) is specific to D-type strains as a group while no proteins is specific to the S-type strains as a group. Because *i*VF846 is based on the S-type ES114 and does contain any of the unique D-type proteins, we tried to understand the differences between the D-type and S-type strains based on the previous annotation of the 194 D-type specific proteins that likely make a difference in co-colonization phenotypes. However, annotating these D-type specific proteins is challenging and a lot of them were “hypothetical” [55]. Next, we searched for functional annotation of the 194 orfs using RefSeq [73], PATRIC [74], and UniProt [75] databases. We found that some of these 194 previously identified D-type specific proteins also exist in S-type strains probably due to errors in high-throughput

genomics comparison and excluded these erroneously categorized proteins from our analysis.

Four D-type specific proteins are likely used for producing specific O-antigens (Table 2.3). From the genomic evidence of the O-antigen related D-type specific genes, D-type and S-type strains might have O-antigens that result in their different motility and co-colonization phenotypes since a previous study showed that an ES114 O-antigen ligase mutant had a motility defect that caused significant delay in colonization and could not compete with the wild type in co-colonization assays [76].

Protein	Annotation in [55]	RefSeq Annotation	PATRIC Annotation (by RAST)	UniProtKB
prot_00668	hexapeptide repeat-containing transferase	acetyltransferase	dTDP-4-amino-4,6-dideoxygalactose transaminase (EC 2.6.1.59)	O-acetyltransferase
prot_00677	hypothetical protein	glycosyltransferase WbuB	capsular polysaccharide synthesis enzyme Cap5L	glycosyltransferase WbuB
prot_00679	hypothetical protein	NAD(P)-dependent oxidoreductase	probable dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)	dTDP-4-dehydrorhamnose reductase
prot_02973	hypothetical protein	*N*-acetyltransferase	Uncharacterized N-acetyltransferase YedL	Acetyltransferase

Table 2.3 D-type Specific Proteins Involved in O-Antigen Biosynthesis.

Three other proteins are likely to be involved in the synthesis of peptidoglycan, a signaling molecule recognized by animals for the presence of bacterial (Table 2.4). In *V. fischeri*, peptidoglycan monomers have been shown to induce the development of squid host tissues and cause the trafficking of host phagocytic hemocytes [77].

Protein	Annotation in [55]	RefSeq Annotation	PATRIC Annotation (by RAST)	UniProtKB
prot_ 00685	glycerol-3-phosphate cytidiltransferase	glycerol-3-phosphate cytidylyltransferase	Glycerol-3-phosphate cytidylyltransferase (EC 2.7.7.39)	Glycerol-3-phosphate cytidylyltransferase
prot_ 00080	penicillin-binding protein 1C	penicillin-binding protein 1C	Penicillin-insensitive transglycosylase (EC 2.4.2.-) & transpeptidase PBP-1C	Penicillin-binding protein 1C
prot_ 00075	hypothetical protein	glutathionylspermidine synthase family protein	Putative acid--amine ligase YgiC	Glutathionylspermidine synthase

Table 2.4 D-type Specific Proteins Involved in Peptidoglycan Synthesis.

Additionally, we gathered hints from annotations of three other proteins about colonization mechanism of the D-type strains (Table 2.5). The annotations of prot_00751 from the previous study and three databases all indicate that this protein binds/processes peptidoglycan. In RefSeq, prot_00751 is shown to have a Lysin Motif (LysM, Pfam PF01476), which binds to various types of peptidoglycan and chitin, and majority of these LysM proteins recognize *N*-acetyl-glucosamine moiety and are peptidoglycan hydrolases [78]. Vgr-3 protein, which is the PATRIC annotation for prot_00751, is also a peptidoglycan binding protein. In *Vibrio cholerae*, VgrG-3 (VC_A0123) acts to degrade peptidoglycan and assists in the delivery of accessory toxins of *V. cholerae* [79]. *V. cholerae* also expresses an anti-toxin encoded immediately downstream (VC_A0124) to avoid self-intoxication [79]. In the D-type strains, an anti-toxin (prot_04197) is encoded as well, although prot_04197 is not directly downstream of the peptidoglycan degradation protein prot_00751. Finally, prot_00073 is unanimously annotated as a phage shock protein A. This protein is known to suppress σ^{54} dependent transcription [80]. *V. fischeri*

σ^{54} has been shown to play a requisite role initiating the colonization by controlling motility and influences biofilm development, nitrogen assimilation, and bioluminescence [81].

Protein	Annotation in [55]	RefSeq Annotation	PATRIC Annotation (by RAST)	UniProtKB
prot_ 00751	cell-wall degradation protein	LysM peptidoglycan-binding domain-containing protein	VgrG-3 protein	putative cell wall degradation protein
prot_ 04197	antitoxin	type II toxin-antitoxin system Phd/YefM family antitoxin	RelB/StbD replicon stabilization protein (antitoxin to RelE/StbE)	Antitoxin
prot_ 00073	phage shock protein A	phage shock protein A	phage shock protein A (IM30), suppresses sigma54-dependent transcription	phage shock protein A

Table 2.5 Three Other D-type Specific Proteins That Are Potentially Related to Vibrio-Squid Interactions.

2.1.4 Incorporating Bioluminescence in *i*VF846

Bioluminescence, a group-dependent behavior regulated partially by quorum sensing in *V. fischeri* is an important component in its metabolism. Although the specific mechanisms that control the light production *lux* operon is yet unknown [82], we incorporated into *i*VF846 genes known to encode all the structural components necessary for bioluminescence [83], i.e., *lux CDABEG*, which are associated with flavin reductase (FMNRx), fatty acid reductase (FAR), and luciferase (LUC) reactions. An additional sink reaction is in *i*VF846 for the photons produced by the luciferase (EX_hv_e) (Figure 2.6 adapted from [83], Table 6). Except the FMNRx reaction that already exists in the *E. coli* *i*JO1366 model, all three other bioluminescence related reactions are unique to *i*VF846. Light production by the bioluminescence pathway is an energy-intensive process in *V.*

fischeri. Flavin reductase reduces flavin mononucleotide (FMN) to its reduced form ($FMNH_2$). Fatty acid reductase reduces a fatty acid (RCOOH) to a long chain aldehyde (RCHO). Luciferase takes in oxygen and oxidizes the RCHO and $FMNH_2$, and releases light. Although the long-chain fatty acid and $FMNH_2$ are recycled in this pathway, reducing power, ATP, and oxygen are consistently consumed.

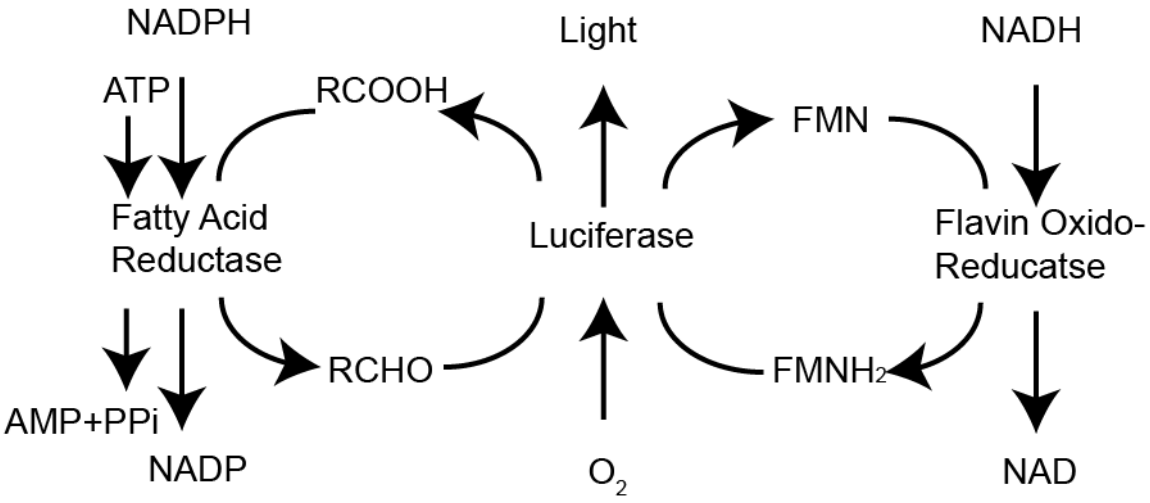


Figure 2.6 Bioluminescence Pathway in *V. fischeri* Involving Fatty Acid Reductase, Luciferase, and Flavin Oxido-Reductase.

Reaction Name	Rxn Abbr.	Equation
Luciferase*	LUC	0.7 atp[c] + o2[c] + ddcald[c] + fmnh2[c] -> 0.3 h2o[c] + 0.7 amp [c] + 1.7 h[c] + 0.7 ppi [c]+ ddca [c] + fmn[c] + 0.1 hv
Fatty acid reductase	FAR	h[c] + ddca[c] + atp[c] + nadph[c] -> ddcald[c] + amp[c] + ppi[c] + nadp[c]
Flavin reductase**	FMNRx	fmn[c] + h[c] + nadh[c] -> fmnh2[c] + nad[c]
Net reaction		o2[c] + atp[c] + nadph[c] + nadh[c] +h[c] -> h2o[c] + amp [c] + ppi[c] + nadp[c] + nad[c] + hv
Luciferase exchange flux	EX_hv_e	hv ->

Table 2.6 Reactions in iVF846 That Are Involved in the Bioluminescence Pathway. (* Different quantum yields and ATP consumption per photon produced vary the stoichiometric coefficients of this reaction. Here the stoichiometric coefficients are calculated for a 10% quantum yield and 7 ATP consumed per photon produced. ** Reaction already exists in *i*JO1366.)

At least as far back as the 1970s scientists have been trying to quantifying the energy cost of bioluminescence and its impact on cell growth by measuring bioluminescence activity and quantum yield, which is the ratio of photons emitted per O_2 molecule consumed [52,84–87]. While bioluminescence activity shows the energy emitted by cells in terms of light per unit time, quantum yield influences the efficiency of bioluminescence, which is reflected by the stoichiometric coefficients of LUC. Using experimentally measured data and our model, we were able to quantitatively assess the impact of bioluminescence on cell growth rate. The measured quantum yield was about 10% in *vitro* but could be better *in vivo* [84,86]. Bioluminescence activity has also measured previously for the *V. fischeri* ES114 strain in 24°C high-osmolarity seawater-tryptone (SWTO) medium or under other similar culturing conditions [52,85]. Under these conditions, the wild type ES114 was very dim and the maximum bioluminescence activity measured was about 3.3 photons/second · cell. Therefore, we estimated the luciferase reaction flux of *V. fischeri* ES114 strain was about 4.4×10^{-5} mmol photons gDW^{-1} h^{-1} (assuming a cell dry weight of 0.45 gDW/OD · L from our dry weight measurement and a concentration of 10^9 cells/mL for a culture around optical density of one). The luciferase reaction flux by ES114 was about 860 times higher (3.7×10^{-2} mmol photons gDW^{-1} h^{-1}) in the squid light organ, and 1500-fold higher (6.6×10^{-2} mmol photons gDW^{-1} h^{-1}) when autoinducer (AI) was added to SWTO medium.

Here we used the model to calculate growth rates of the wild type ES114 strain by varying the levels of luciferase flux and quantum yield (Figure 7). At a fixed level of quantum yield (10%, 20%, 50%, or 100%), the growth rate decreases as the luciferase reaction flux becomes higher. The decrease in growth rate becomes faster at a lower

quantum yield or a lower oxygen uptake rate. From our calculated luciferase reaction flux values of ES114 in SWTO medium without AI and in squid (4.4×10^{-5} and 3.7×10^{-2} mmol photons gDW^{-1} h^{-1} respectively, which are indicated by the yellow and purple vertical lines, Figure 2.7), we estimated that, when quantum yield is only 10%, the growth rate is about 13% lower in squid with a low oxygen uptake rate, 9% lower with a medium oxygen uptake rate, and 5% lower with a high oxygen uptake rate. When quantum yields is 100%, the cell growth rate is about 5%, 3%, and 2% lower with low, medium, and high oxygen uptake rates respectively. Regardless of the levels of quantum yield and oxygen uptake rates, our model indicates bioluminescence creates a noticeable energy cost to *V. fischeri*, which is consistent with experimental evidence that the induction of luminescence in the wild type *V. fischeri* ES114 strain slowed growth relative to dark mutants [85]. Notice that our estimation does not include the cost of protein synthesis, which indicates the true cost is higher than we what we estimated. Although bioluminescence resulted in a significant amount of energy consumption, the wild type achieved about four fold higher populations than the dark Δ*luxCDABEG* mutants in the light organ of the squid [85]. This shows that the interactions *V. fischeri* established with the squid in light organ overturned the disadvantages of bioluminescence in terms of energy costs into colonization advantages.

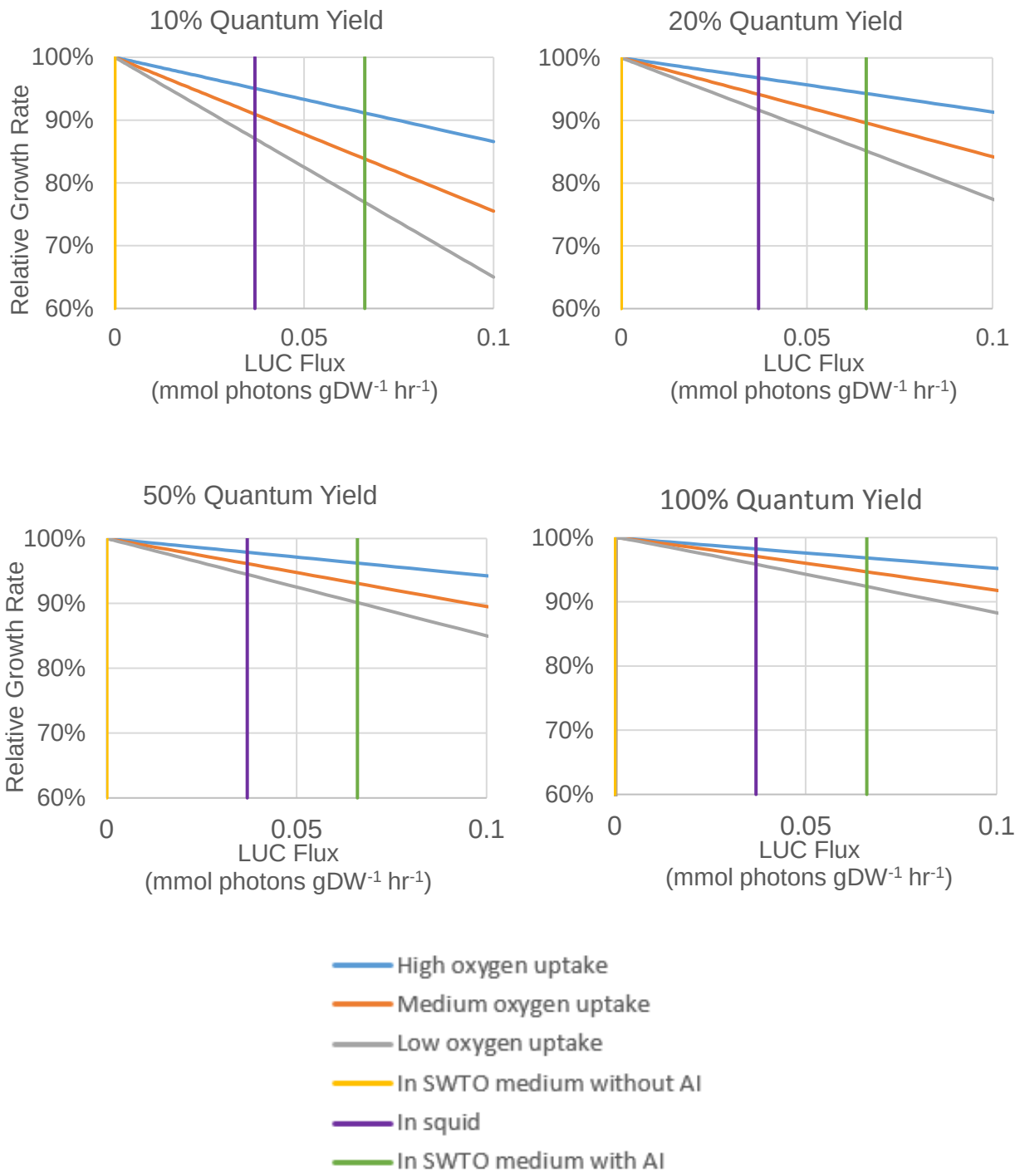


Figure 2.7 Relative Growth Rate of *V. fischeri* vs. Luciferase Reaction Flux. The blue, orange, and gray lines show the growth rates at different values of LUC flux normalized by the maximum growth rate at zero luciferase reaction flux. The vertical yellow, purple, and green lines indicated the experimentally-derived luciferase flux values in SWTO medium without AI, in squid, and in SWTO medium with AI respectively.

2.2 Conclusions

In this study, we presented a genome-scale metabolic reconstruction of *V. fischeri* ES114, *i*VF846, which constitutes of our current knowledge of metabolic genes and reactions in this bacterium. We measured biomass composition, cell dry weight, and growth rates of ES114 in laboratory culturing environment for parameterizing the model, performed high-throughput growth-phenotyping experiments and incorporated growth phenotype and gene essentiality data for model curation and validation. During this process, we were able to find genes for four non-GPR reactions in the draft model using a model-enabled gene search approach [58]. The final reconstruction provides predictions that match well with experimental results with an 87.9% accuracy in predictions of growth phenotypes on sole carbon sources and a 91.6% accuracy in predictions of essential genes. Only seven gap-filled non-GPR reactions exist in the final model, five of which are transport reactions that usually occur in GENREs due to poor annotation of transporters.

After *i*VF846 was curated, we demonstrated its ability to locate unique features in *V. fischeri* metabolism that is potentially associated with squid colonization, quorum sensing, and bioluminescence. We first compared *i*VF846 with the *E. coli i*JO1366 genome-scale metabolic reconstruction and found a large number of catabolic reactions for carbon sources and other nutrients unique to each organism. These differences might arise from the nutrients provided by their different hosts. We also noticed a large number of *V. fischeri* or *E. coli* specific reactions involving in the biosynthesis of membranes/lipids/LPS, representing the intrinsic difference in their biomass composition, many of which might influence the Vibrio-squid recognition and interactions. Through the comparison between *i*VF846 and *i*JO1366, we noticed a *V. fischeri* histidine catabolic pathway which can be down-regulated by *N*-acetyl-D-glucosamine (**Chapter 3.1.2**), a

sugar key to the development and maintenance of its symbiosis relationship with the squid. In addition, *V. fischeri* has a unique high oxygen affinity cytochrome (CcoNOQP) that can be down-regulated by *N*-acetyl-D-glucosamine. We suspected that CcoNOQP can serve as a control node by the squid for obtaining optimal bioluminescence from *V. fischeri* at the right time.

Comparing *i*VF846 to the core genome of 13 *V. fischeri* strains, we found that ES114 and the 12 other *V. fischeri* strains have very similar metabolic capabilities. The metabolic functions unique to ES114 (at least the metabolic functions that are well characterized and included in our model) are mostly related to the transport of nutrients, oxidative phosphorylation, and LPS biosynthesis. Therefore it is possible that ES114 utilizes substrates and respiratory pathways that are slightly different from other *V. fischeri* strains under certain environmental conditions although very few of these metabolic functions might be involved in squid colonization.

We found consistent and informative functional annotation for ten proteins that might determine the colonization phenotypes ("dominant" or "sharing") of the 13 *V. fischeri* strains in pairwise competition assays. All these ten proteins are related to the processing of o-antigen, peptidoglycan, and outside toxins, but are not related to substrate utilization or respiration. Although we are able to understand the functions of the ten proteins, the majority of them are not well annotated in databases, and extensive future functional annotation work is required for a better understanding of *V. fischeri* strain-specific colonization phenotypes.

Finally, using the bioluminescence reactions in *i*VF846, we calculated the impact of bioluminescence on cell growth rates at various levels of oxygen uptake and quantum

yield and observed a much slower *in silico* growth rate of *V. fischeri* in squid than in laboratory medium. This difference in growth rate is more significant when levels of oxygen uptake and quantum yield are low.

From this study, we showed that our *i*VF846 genome-scale metabolic network for *V. fischeri* ES114 serves as a useful tool to understand *V. fischeri* metabolism. It provides a genome-scale view of the diverse metabolic capabilities of *V. fischeri* have for adapting to free-living and host-associated life styles. We expect it to guide researchers to make further discoveries in *V. fischeri* metabolism and its relationship with the squid. Future discoveries will be greatly facilitated by continued improvement of *i*VF846 because some important features of *V. fischeri* metabolism, e.g. Na^+ translocating reductases, are not included in *i*VF846 yet, due to a lack of information in the detailed mechanisms and reaction stoichiometry of these features. The *i*VF846 also did not include unannotated hypothetical genes or genes that are poorly annotated although many of them have been found to play a role in Vibrio-squid interactions, quorum sensing, and bioluminescence.

Chapter 3

Transcriptomic Analysis for Understanding *Vibrio fischeri* Metabolism

Some material in this chapter has been adapted from:

> Thompson LR, Nikolakakis K, Pan S, Reed J, Knight R, Ruby EG: **Transcriptional characterization of *Vibrio fischeri* during colonization of juvenile Euprymna scolopes**. *Environ. Microbiol.* 2017, **19**:1845–1856.

Transcriptomic experiments are commonly performed by biologists to explore transcriptional responses to perturbations in biological systems. However, processing large transcriptomic datasets and discovering patterns in them could be daunting and laborious. In addition to the popular functional enrichment analysis, we used genome-scale metabolic modeling that takes advantages of the mass-balanced reaction network of *V. fischeri* and systematically analyzed the two datasets from our collaborators in the Squid and Vibrio labs. Such approach allowed us to make novel metabolic discoveries about *V. fischeri* metabolism, which might not be obvious to biologists manually going through the list of differentially expressed genes.

The first RNA-Seq dataset that we analyzed represented host-associated *V. fischeri* cells from colonized juvenile squid and under laboratory and simulated marine planktonic conditions. The picture that arises of the establishment of the squid-Vibrio symbiosis is one of tightly regulated developmental events within the host, and a well-adapted suite of tools for host colonization on the part of the bacterial symbionts. Prior studies have examined transcriptional patterns within this system in mature host animals, and within the host light organ during initial contact with bacterial symbionts; however,

transcriptomics data on the bacteria in the juvenile light organ has thus far been missing because there are $<10^6$ symbionts per light organ during this early stage of development, below the input biomass limit of standard RNA-Seq transcriptomics. In light of the strong and early response by the squid host to the initial presence of the symbiont [88], we asked what changes occur within the symbionts as they make the transition from the bacterioplankton to the host, a habitat switch similar to that made by most pathogenic Vibrio species. We performed RNA-Seq experiments on *V. fischeri* from three environments—(*i*) cells growing in rich medium, (*ii*) cells incubated in seawater and (*iii*) cells collected immediately after venting from the squid host—to identify transcriptional signals that could be directly associated with the early stages of host colonization. Flux coupling analysis based on this dataset was used to nutrients sources provided to *V. fischeri* by juvenile squids.

The other RNA-Seq dataset represented *V. fischeri* cells catabolizing chitins and under no chitin catabolism conditions. Chitins derived from the squid play an important role in the symbiotic relationship between squid and *V. fischeri*. One of the roles of chitins in this relationship is to change the physiology of *V. fischeri* by chitin catabolism so *V. fischeri* can acidify the adult light organ, which increases the level of oxygen available for bioluminescence at night [72,89]. A previous study revealed details about how sugar transport, acetate switch, and cytochrome oxidase activity are regulated by chitin catabolism. Here we collected a complete transcriptomic RNA sequencing dataset from the growth of a *V. fischeri* wild type and a chitin transporter mutant Δ*nagE* with and without the chitin, i.e. , *N*-acetyl glucosamine (GlcNAc), and identified up- and down-regulation of

pathways at systems level using functional enrichment analysis and genome-scale metabolic modeling.

3.1 Results and Discussion

3.1.1 Nutrient Sources of *V. fischeri* in Juvenile Squids

To investigate which nutrients might be used by the symbionts of juvenile squids, we applied flux coupling analysis [13] using a *V. fischeri* genome-scale metabolic model. This analysis can identify metabolic genes that are linked to the utilization of specific carbohydrates, such that flux through the reactions catalyzed by these genes will indicate that a particular substrate is being used by the bacteria. This flux coupling analysis identified *N*-acetylneuraminic acid (NANA) as a possible carbon source for *V. fischeri* in juvenile animals based on an apparent upregulation of several genes involved in NANA uptake and utilization. Independent support for NANA as a symbiont substrate comes from a high-throughput insertion sequencing study [54] predicting genes encoding the putative transporter and lyase of NANA (nanT and nanA) to be colonization factors. In addition, NANA is a component of squid mucus, and a chemoattractant of *V. fischeri* [68,90]. The analysis also suggested glycerophosphodiester as a potential carbon source for *V. fischeri* in juvenile animals based on the upregulation in the symbionts of *glpQ*. In the analysis, *glpQ* is the only gene specifically connected to the utilization of glycerophosphodiesters, because other genes involved in glycerol and G3P catabolism also play a role in the use of other nutrients. Overall, these results support the conclusion that nutrients such as NANA and glycerophosphodiester may be used as carbon sources by symbionts in the juvenile light organ. Future metabolomics studies aimed at detecting these and other potential carbon sources in the host will help refine comparisons between our flux coupling analysis and transcriptomic results.

3.1.2 Differentially Expressed Metabolic Gene Sets upon Chitin Addition

We performed gene set enrichment analysis (GSEA) according to an established protocol [91] and obtained an understanding of the genome-scale metabolic pathway changes instead of single gene changes imposed by the addition of GlcNAc on *V. fischeri* using three pairwise comparisons (GlcNAc EIIBC transporter mutant Δ*nagE* vs. Δ*nagE* + GlcNAc, wild type vs. wild type + GlcNAc, and Δ*nagE* + GlcNAc vs. wild type + GlcNAc). Because growth of GlcNAc requires *nagE* [72] and GlcNAc is unlikely to get metabolized efficiently by Δ*nagE*, it was expected that GSEA showed that very few genes were differentially expressed in the case of Δ*nagE* vs. Δ*nagE* + GlcNAc, and the only two gene sets with a false discovery rate (FDR) less than 0.25 are tRNA charging and tetrapyrrole biosynthesis (from glutamate), which were both differentially expressed in the other two comparisons. The comparisons between wild type vs. wild type + GlcNAc, and Δ*nagE* + GlcNAc vs. wild type + GlcNAc, show similar highly enriched gene sets (Figure 3.1).

Figure 3.1 Top 50 Differentially Expressed Gene Sets. The absolute values shown for each gene set is the log10 (FDR). The signs (- and +) indicate down- and up-regulation respectively. There are two pairs of comparisons, i.e., the wild type vs. wild type + GlcNAc and ΔnagE + GlcNAc vs. wild type + GlcNAc conditions, and a total of 264 gene sets.

Specifically, GSEA showed that upon GlcNAc addition the central metabolism was rewired for the production of acetate from GlcNAc (Figure 3.2). First, glycolysis was up-regulated and TCA cycle was down-regulated. Meanwhile, aerobic respiration through the ccb3-type cytochrome oxidase CcoNOQP was turned down. This is consistent with the results from a previous study where Δ*ccoNOQP* mutant that first suffered from poorer growth rate relative to wild type restored growth rate when GlcNAc was added [72]. Among the top differentially expressed pathways also includes the up-regulated fermentation pathway of pyruvate to acetate through the acetyl transferase (Pta) and acetate kinase (AckA). Combining observations in these above pathways, we suspect that the increased glycolysis flux upon addition of GlcNAc was first directed to pyruvate, and then to acetate through fermentation (Figure 3.2). In addition to the central metabolism changes caused by GlcNAc utilization, GSEA indicated that GlcNAc was used as a preferred PTS carbohydrate and repressed the catabolism of non-PTS sugars —including glycerol, glycerophosphodiesters, *N*-acetyl neuraminate, and *N*-acetyl mannosamine—and the PTS sugar galactose.

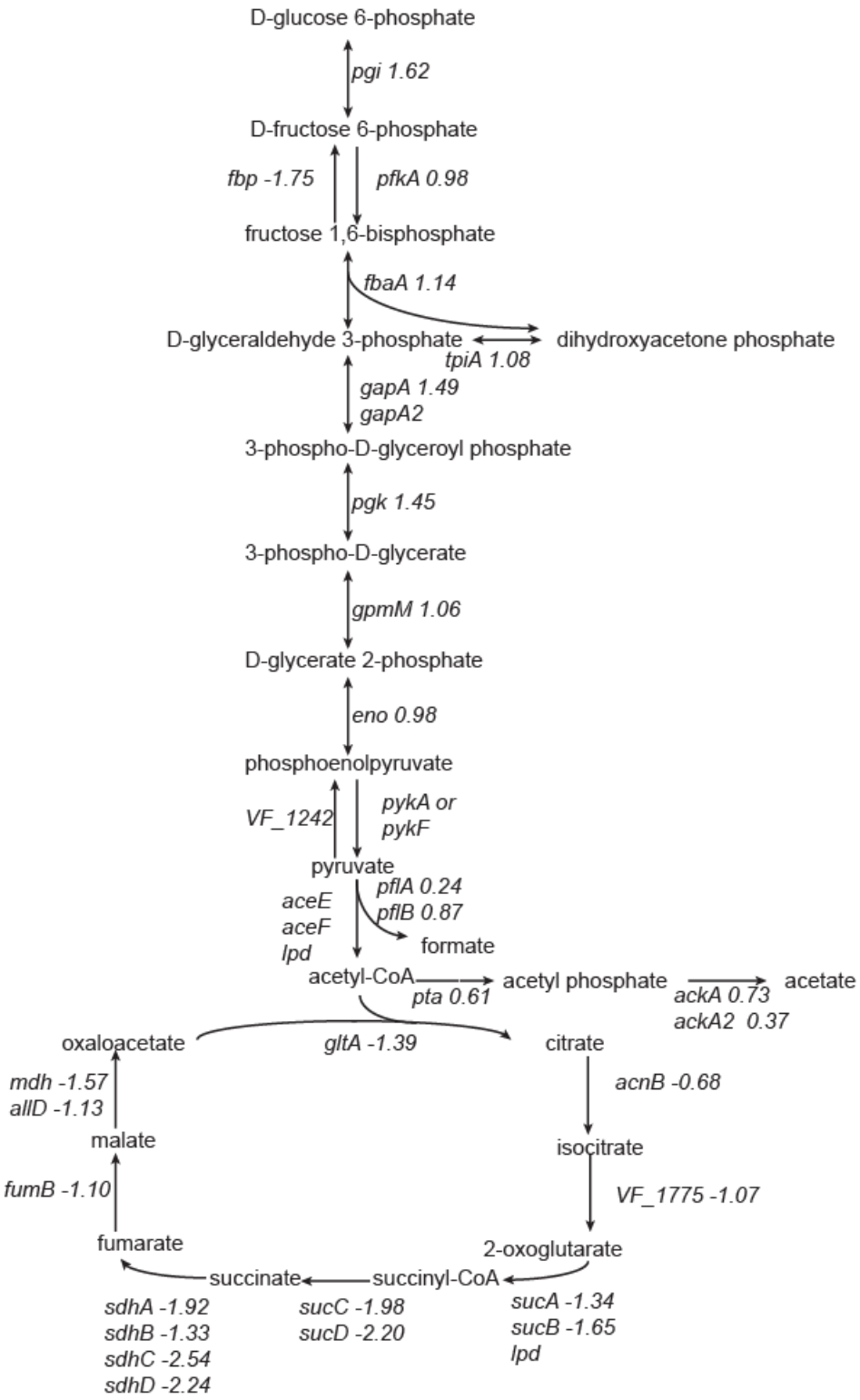


Figure 3.2 Central Metabolic Pathways Regulated by GlcNAc. Genes were labeled with log2 fold changes (FCs) of their expression levels under the conditions of Δ*nagE* + GlcNAc vs. wild type + GlcNAc. A positive number indicate a higher expression level in wild type + GlcNAc.

Along with the redirected carbon flow in central metabolism, another impact of GlcNAc catabolism was the changes in pathways related to nitrogen metabolism, which include *i*) amino acid metabolism of proline, histidine, arginine (Figure 3.3), and glycine (Figure 3.4), and *ii*) purine metabolism (Figure 3.4). The level of glutamate, a focal point of nitrogen assimilation, was likely influenced by the down-regulation of ammonium- and glutamate-producing histidine and arginine degradation pathways upon GlcNAc addition (Figure 3.3). The degradation of proline also produces glutamate. Although proline degradation pathway was excluded in GSEA because this pathway includes less than three genes, both genes in the proline degradation pathway, i.e., *putA* and VF_A0830, were down-regulated (-1.34 and -1.21 fold respectively, Figure 3.3).

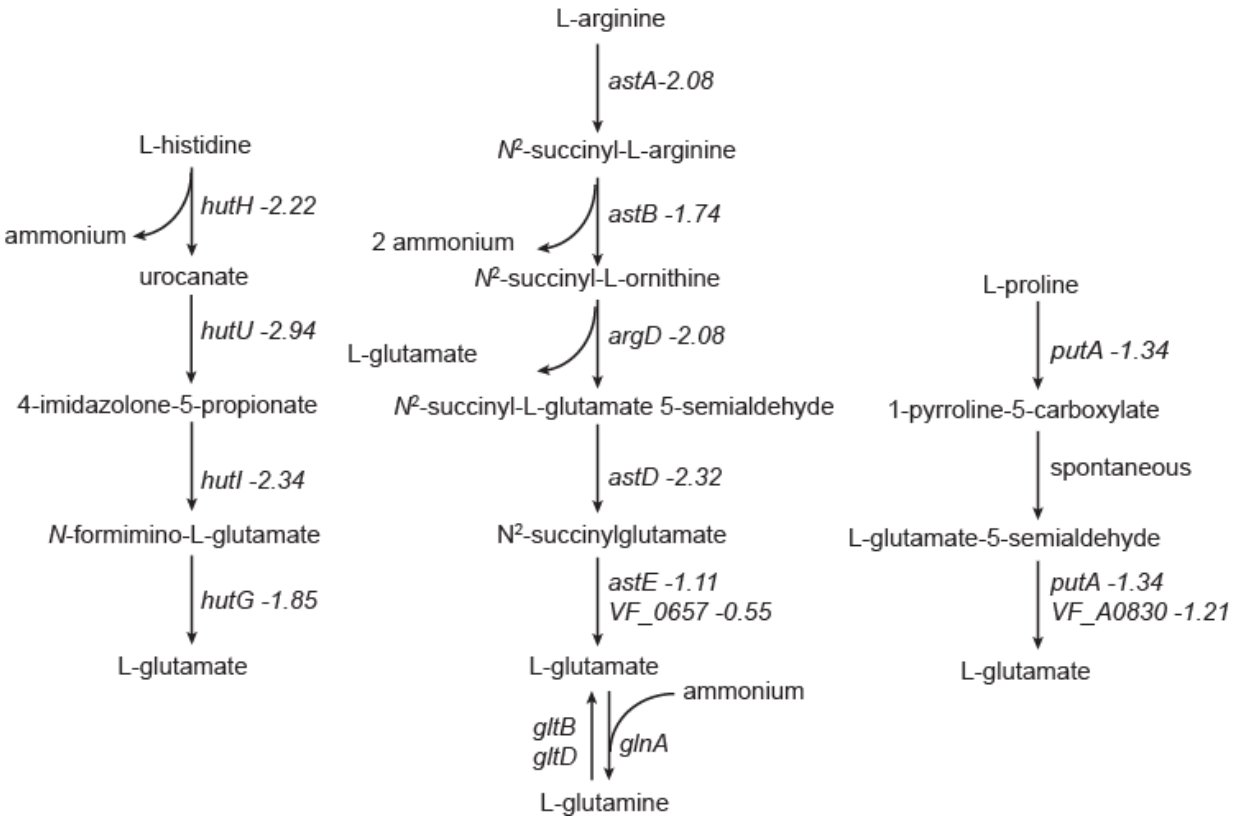


Figure 3.3 Arginine, Histidine, and Proline Catabolism Pathways Regulated by GlcNAc. The specific arginine catabolism pathway involved is the arginine succinyltransferase (AST) pathway. Genes were labeled with log2 FCs of their expression levels under the conditions of Δ*nagE* + GlcNAc vs. wild type + GlcNAc. A positive number indicate a higher expression level in wild type + GlcNAc.

Opposite of the down-regulation of the abovementioned ammonium- and glutamate-producing amino acid metabolism pathways, the biosynthesis of the purine intermediate inosine monophosphate (IMP), which takes in ammonium from glutamine and glycine, was up-regulated. At the same time, glycine cleavage pathway was significantly down-regulated, possibly accumulating glycine to supply the up-regulated IMP biosynthesis. (Figure 3.4).

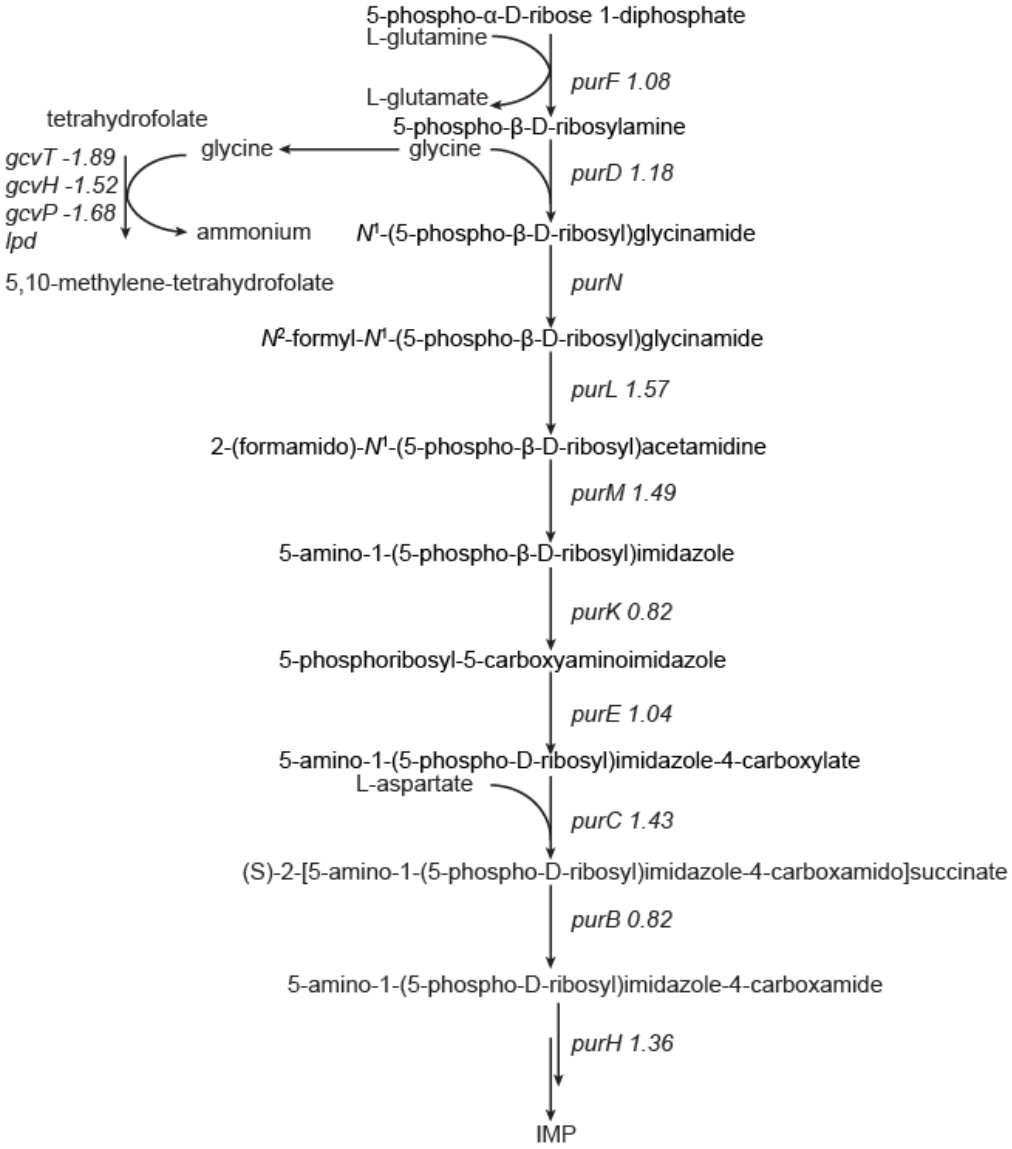


Figure 3.4 IMP Biosynthesis and Glycine Cleavage Pathway Regulated by GlcNAc. Genes were labeled with log2 FCs of their expression levels under the conditions of Δ*nagE* + GlcNAc vs. wild type + GlcNAc. A positive number indicate a higher expression level in wild type + GlcNAc.

3.1.2 Acetate and Ammonium Secretion upon Growth on Different Sugars

As shown earlier, GSEA revealed that GlcNAc likely regulated the *V. fischeri* central metabolism—to allow the acetate-producing fermentation of GlcNAc—and nitrogen metabolism. We further assessed whether there exist physiological benefits of GlcNAc catabolism in *V. fischeri* by simulating acetate and ammonium secretion using the genome-scale metabolic model of *V. fischeri*, *i*VF846.

First, we simulated—under an anaerobic or medium level of oxygen uptake and a varying sugar uptake rate condition—the acetate secretion from the catabolism of five different sugars, i.e., GlcNAc, chitobiose, glucosamine, glucose, and glycerol, which is representative of PTS and non-PTS sugars and amino and non-amino sugars (Figure 3.5). Among the sugars we tested using modeling, GlcNAc and chitobiose are the most acidogenic and secrete acetate the fastest at all levels of oxygen. And as oxygen becomes limited, more acetate is secreted via Pta and AckA, a fermentation pathway that was significantly up-regulated in our experiment where GlcNAc was added to well-oxygenated *V. fischeri* cells with vigorous shaking. This shows that despite the sufficient oxygen supply, GlcNAc enabled *V. fischeri* to operate in an acetate-producing, oxygen-limiting fermentation mode. In *E. coli*, the protein level of Pta is decreased under low pH conditions [92]. In contrast, in our experiment, *V. fischeri* up-regulated the acetate-producing Pta-AckA pathway upon GlcNAc addition which potentially allowed *V. fischeri* to produce even more acetate under a low pH condition. Additionally, the expressed transcripts of acetyl-coA synthetase (*acs*), which is responsible for acetate uptake, and its transcriptional activator (*ainS*) in wild type, decreased significantly upon GlcNAc addition (log2 FC = -1.14, FDR = 3.94E-10 for *acs*; log2 FC = -1.59, FDR = 2.02E-23 for *ainS*, consistent with data shown in [72]). Therefore, in *V. fischeri*, GlcNAc not only can

up-regulate the acetate-producing Pta-AckA fermentation pathway but also can inhibit the uptake of secreted acetate by repressing *acs* transcription. When *V. fischeri* is inside the squid light organ, such a process can accumulate acetate and acidify the environment for bioluminescence, which makes GlcNAc an excellent nutrient choice by the squid at night.

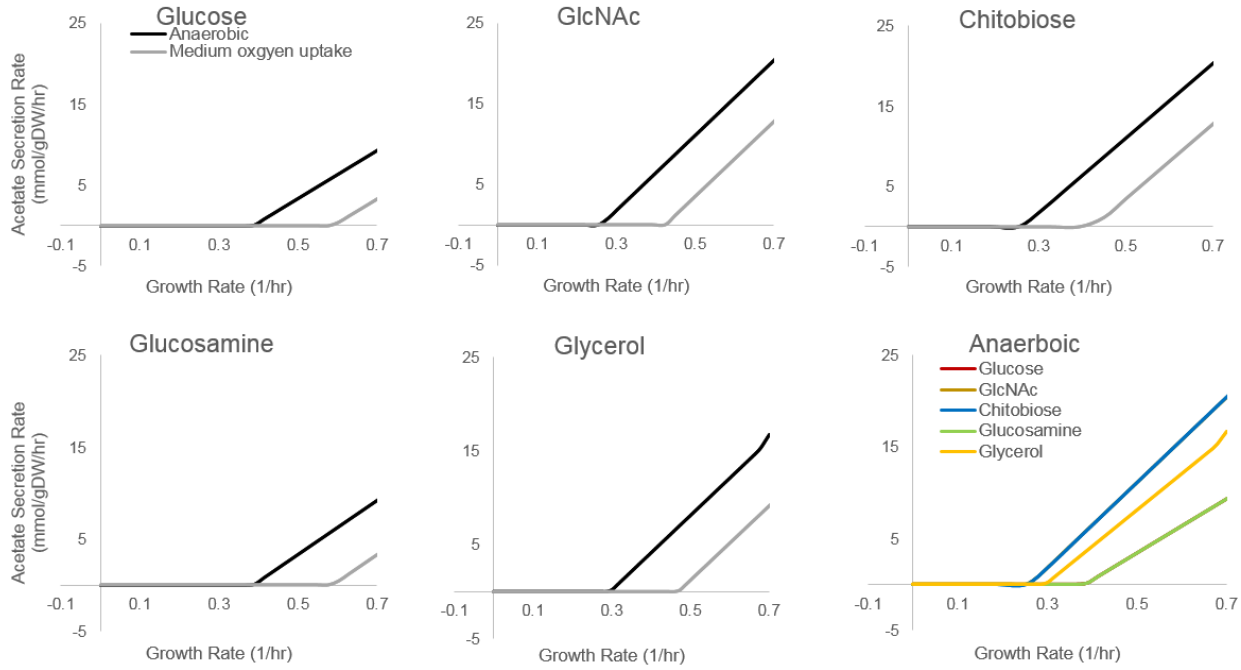


Figure 3.5 Acetate Secretion Profiles. In the simulations, glucose, GlcNAc, chitobiose, glucosamine, and glycerol were tested under an anaerobic or medium level oxygen uptake condition. In the bottom right panel of this figure, the overlaying lines for GlcNAc and chitobiose conditions are the steepest and the overlaying lines for glucose and glucosamine conditions are the least steep.

Finally, we simulated ammonium secretion from the catabolism of the same five sugars assessed earlier (Figure 3.6). At a low growth rate, ammonium secretion can vary at the maximum growth rate because other nitrogen-containing metabolites instead of ammonium can be secreted. At a high growth rate, ammonium becomes the preferred metabolite for secreting excess nitrogen. We observed a general trend similar to acetate secretion where ammonium secretion is higher when oxygen becomes limited. Comparing to glucose and glycerol, the amino sugars—GlcNAc, chitobiose, and

glucosamine—secret ammonium more rapidly when growth rate increases. It is possible upon GlcNAc addition the observed up-regulation of IMP biosynthesis, which consumes glycine and glutamate, and down-regulation of proline, histidine, and arginine catabolism, which produces ammonium and glutamate, provides a strategy for better *V. fischeri* growth.

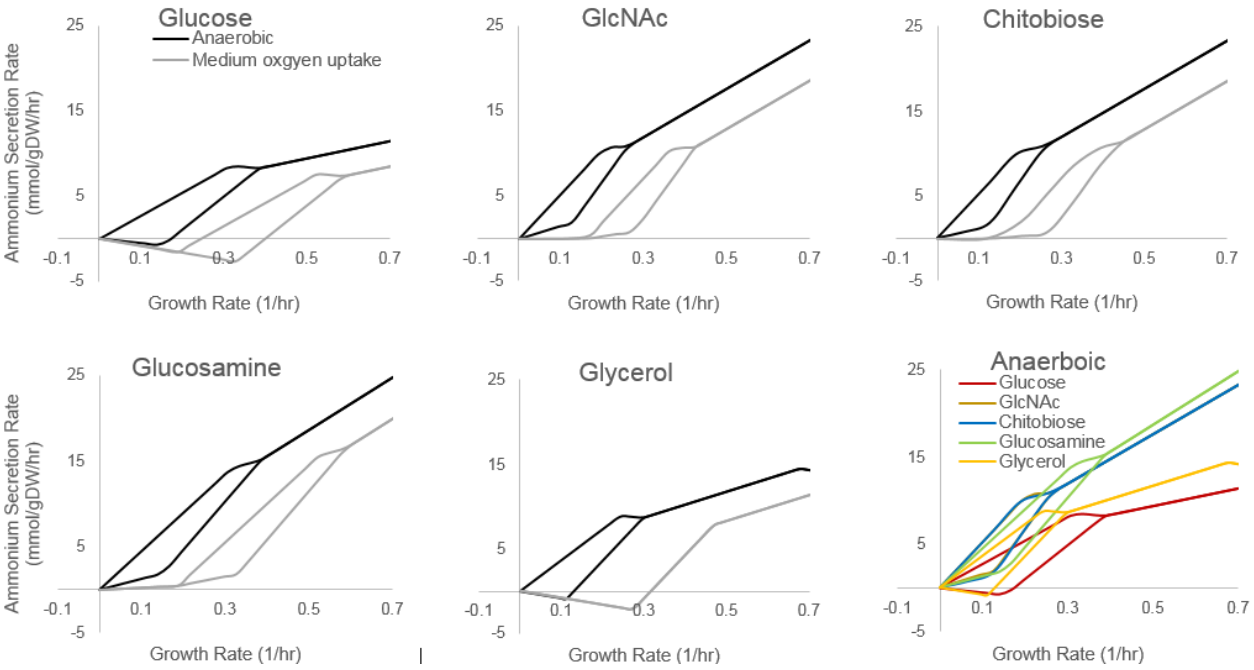


Figure 3.6 Ammonium Secretion Profiles. In the simulations, glucose, GlcNAc, chitobiose, glucosamine, and glycerol were tested under an anaerobic or medium level oxygen uptake condition. At lower growth rates, ammonium secretion can vary at the maximum growth rate. The lower and upper limits of this range are shown in this figure, and any ammonium secretion rate within the limits is feasible. In the bottom right panel of this figure, the lines for GlcNAc and chitobiose overlay towards higher growth rates.

3.2 Conclusions

We demonstrated the effectiveness of our approaches for analyzing large transcriptomic datasets by revealing two nutrient sources of *V. fischeri* in juvenile squids and three groups of gene sets with significant changes in gene expression levels upon GlcNAc addition, i.e., *i*) increased acetate-producing GlcNAc fermentation, *ii*) decreased catabolism of arginine, histidine, and proline, and *iii*) decreased glycine cleavage and increased purine biosynthesis. The latter two groups of gene sets are both related to nitrogen metabolism.

Chapter 4

Model-Enabled Gene Search (MEGS) and Its Applications

Some material in this chapter has been adapted from:

> Pan S, Nikolakakis K, Adamczyk PA, Pan M, Ruby EG, Reed JL: **Model-enabled gene search (MEGS) allows fast and direct discovery of enzymatic and transport gene functions in the marine bacterium *Vibrio fischeri***. *J Biol Chem* 2017, **292**:10250–10261.

The development of next-generation sequencing technologies has generated thousands of genome sequences. These are primarily annotated by a combination of bioinformatics methods which are both they are fast and can be applied genome-wide. Homology-based bioinformatics methods (e.g., BLAST) assume similar sequences share similar functions. Structure-based methods [93,94] and genomic context-based methods (e.g., conserved operon, gene fusions, gene co-occurrence across genomes [95–97]) can be utilized to infer functions that are difficult to annotate using BLAST alone. These bioinformatics methods are often used in conjunction with high-throughput experimental data (including gene expression, protein-protein interactions, mass-spectrometry, RNAi, and mutant fitness) to suggest gene functions based on connections between genes with known and unknown functions. For example, recent studies have used correlations in transposon mutant-fitness scores across multiple experimental conditions to improve genome annotations [98]. Despite the power of these bioinformatics methods, and the increasing availability of high-throughput data, 40 - 60% of newly sequenced genes still lack assigned functions [99–101]. In addition, although bioinformatics methods can quickly predict specific gene functions, biochemical characterization must be still performed

separately to validate those predictions. In fact, a majority of the gene functions assigned have no experimental evidence. For example, as of August 2016 only ~27% of the entries in the UniProtKB Swiss-Prot knowledgebase contain experimental evidence at the protein or transcript level [102]. Direct experimental testing of gene functions that are proposed bioinformatically or based on high-throughput experiments is needed to reduce the high rate of incomplete and incorrect annotations [100,103,104]. Such assessment is important because functions incorrectly assigned to gene sequences enter databases that are subsequently used to assign functions to new sequences. As a result, errors that are hard both to identify and to correct will propagate.

Consequently, it is crucial to develop approaches that quickly identify missing and/or erroneously assigned gene functions, and provide fast and direct experimental validation for the correct function. Such goals can be achieved by combining genome-scale metabolic modeling with experimental techniques. Genome-scale metabolic models are developed primarily based on genome annotations obtained from bioinformatics tools. Current model-based algorithms, including GapFill, SMILEY, and GrowMatch, can use cell-culture data to pinpoint knowledge gaps caused by missing or incorrectly called gene functions; however, these algorithms cannot identify candidate genes for these functions [59,105,106]. More recent algorithms including PHiller-GC, Model SEED, ADOMETA, and MIRAGE identify missing metabolic reactions and candidate genes that might be responsible for catalyzing them [31,107–109], but these algorithms require additional data such as annotated sequences from other organisms and/or expensive gene-expression datasets that might not be available. Importantly, all of these current model-based approaches still do not provide direct experimental validation of the candidate gene's

function. Here, we propose a high-throughput model-enabled gene search (MEGS) method that rapidly identifies functions for unannotated or misannotated genes (**Chapter 4.1.1**). The metabolic modeling procedures in MEGS quickly generate a list of missing or erroneous functions in genome annotations derived from bioinformatics tools, and design functional selection experiments (experiments where only strains that gain an essential function from a genomic library are able to grow) to select for genes with these functions. Subsequent functional selection experiments identify the responsible gene(s) from a genomic library, and provide fast and direct experimental evidence for the gene's function. In contrast to metagenomic functional selections, which have been used to identify RubisCO, DNA polymerase, and antibiotic- resistance genes [110–112], MEGS' functional selections are based on knowledge gaps identified by metabolic modeling. By using genomic functional selections, MEGS does not rely on sequence similarity or genomic context to find gene functions and, as such, can be used to discover functions for previously uncharacterized groups of genes. As such, MEGS complements existing bioinformatics tools to improve genome annotations. Here, we demonstrated a detailed proof-of-concept work in finding *V. fischeri* genes (**Chapter 4.1.2),** and also demonstrated strategies extending MEGS approach for searching orphan enzymes (**Chapter 4.1.3**).

4.1 Results and Discussion

4.1.1 Overview of MEGS

MEGS involves three steps that combine metabolic modeling and experimentation (Figure 4.1). First, a genome-scale metabolic model of an organism of interest is developed, and physiological experiments are performed to validate the model. Computational tools [59,106] are then used to pinpoint metabolic function(s) that are missing from the model, but are needed to resolve model-data discrepancies. The

physiological experiments suggest these missing functions occur, but the discrepancies between model predictions and experiments indicate that the functions are absent from current genome annotations. For example, the *V. fischeri* model originally lacked genes involved in catabolism of both D-xylose and mannitol; however, only genes involved in mannitol catabolism were identified by the model as missing because *V. fischeri* grew on mannitol (but not D-xylose) as a sole carbon source. Second, a recipient strain (derived from a well-characterized organism, e.g., *Escherichia coli*) and selective medium (e.g., a minimal medium supplemented with a single carbon source) are designed such that the recipient strain can only grow in the selective medium if the gene(s) presumptively encoding the missing metabolic function (from the organism of interest) is transferred to the recipient strain. Such pairs of recipient strains and selection media can be computationally designed for a reaction of interest using a forced coupling algorithm [14]. Third, a genomic functional selection experiment is performed to locate the gene(s) in the genome that are responsible for the missing metabolic function and provide direct evidence for the gene's function. For this last step a genomic library of the organism of interest is created by inserting random genomic DNA fragments into plasmids. This plasmid library is then transformed into the recipient strain. Gene(s) responsible for the missing metabolic function can then be identified by sequencing plasmids that complement growth of recipient strains in the selective medium. The discovered genes can then be further characterized and added to the model to improve predictions. As more experimental data are generated, any new model-data discrepancies that arise can be used to drive additional MEGS cycles.

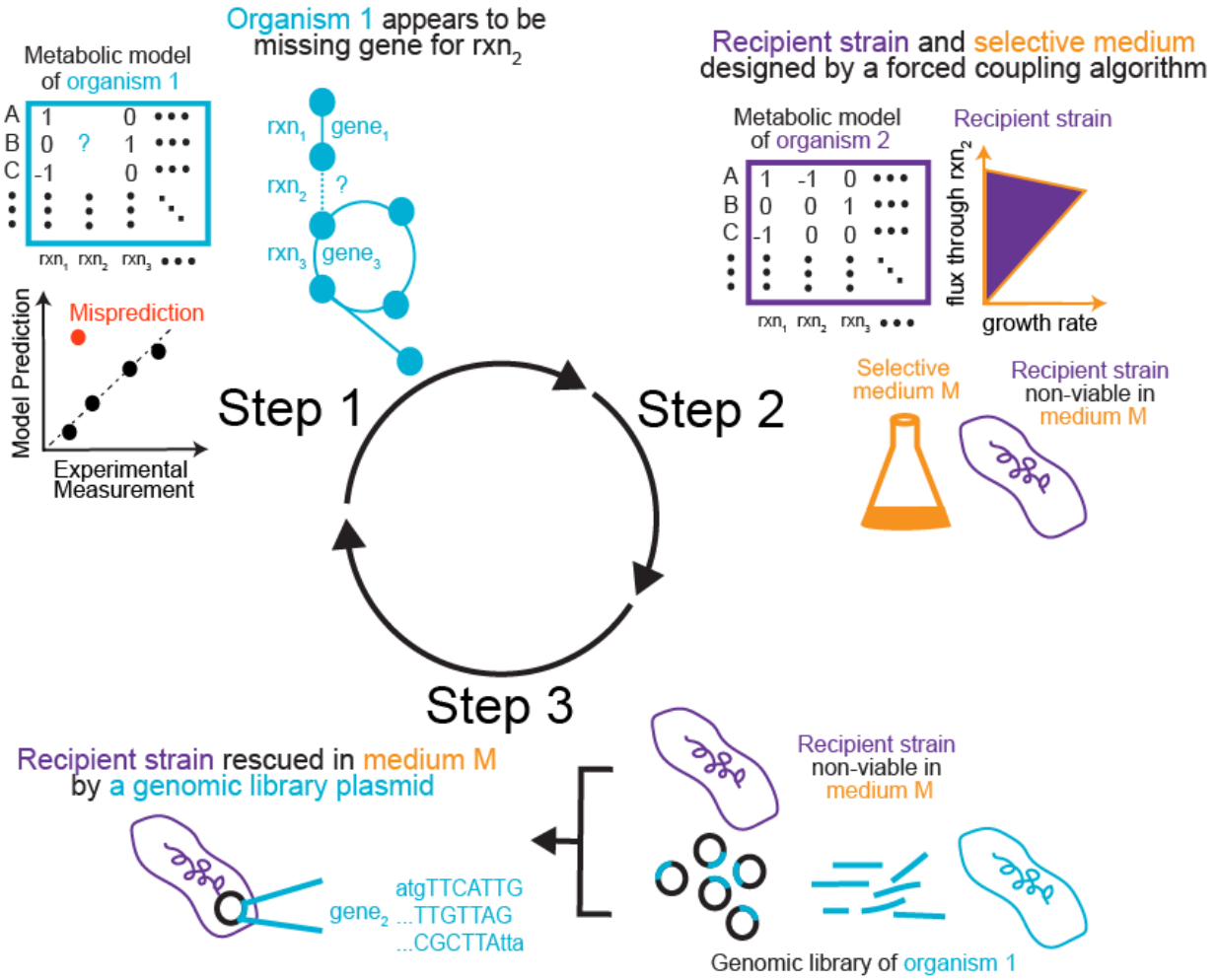


Figure 4.1 Overview of MEGS. In the first step, a metabolic model for an organism of interest is constructed. Target reactions, which are either missing from the model or assigned to the wrong genes, are inferred from discrepancies between model predictions and experimental measurements. A recipient strain (derived from a well-characterized organism, e.g., *E. coli*) and selective medium are then designed for each target reaction. Such recipient strains can only grow in the selective medium if they acquire heterologous enzymes that catalyze the target reaction. Finally, a genomic library is created and the recipient strain is used to select for genes capable of catalyzing the target reactions, since such genes will enable growth of the recipient strain in the selective medium. The discovered genes can then be further characterized, and added to the model to improve predictions.

4.1.2 Discovering *Vibrio fischeri* Gene Functions Using MEGS

In this work, we applied MEGS to discover and characterize several enzyme- and transporter-encoding genes of *V. fischeri* ES114, which is a bioluminescent marine bacterium that forms a symbiotic relationship in the light-emitting organ of the Hawaiian

bobtail squid, *Euprymna scolopes* [113]. Its metabolic capabilities are representative of many other marine bacteria, including both beneficial and pathogenic members of the *Vibrio* genus, and are thus of particular interest [114]. From the KEGG-annotated genome we reconstructed a genome-scale metabolic model for *V. fischeri* ES114, named *i*VF846. Reactions and metabolites from an *E. coli* model, iJO1366, [115] were transferred into the draft model of iVF846 when orthologs to *E. coli* metabolic genes were found in *V. fischeri*. The draft iVF846 model was then curated based on (*i*) data and information reported in the literature, and (*ii*) new growth-phenotyping experiments using Biolog plates, a method for individually testing the ability to metabolize 96 different carbon sources using a microtiter-dish format. To facilitate model curation, a modified version of the SMILEY [59] algorithm (Appendix III Methods) was used to identify missing enzymatic or transport reactions that, if added to the model, would resolve discrepancies between model predictions and experimental growth phenotypes of *V. fischeri* ES114 wild type and mutants. This analysis identified that *V. fischeri* was missing an annotated aspartate 1-decarboxylase (encoded by *panD* in *E. coli*), which caused the draft model to predict no growth either in LBS or in a *V. fischeri* defined minimal medium (DMM) (Figure 2.2a). The growth-phenotyping experiments were performed to identify sole carbon sources that support growth of *V. fischeri* (Appendix III Methods). These results were compared to model-predicted sole carbon sources using flux-balance analysis (FBA) [116], and discrepancies were found for mannitol and *N*-acetylneuraminate. The modified SMILEY algorithm predicted that mannitol-1-phosphate 5-dehydrogenase (encoded by *mtlD* in *E. coli*) and an *N*-acetylneuraminate transporter (encoded by *nanT* in *E. coli*) were missing from the draft model (Figure 2.2b & c). Finally, FBA was used to predict essential *V.*

fischeri genes in LBS medium, and gene essentiality predictions were compared to a recent transposon insertion study [54]. One false-positive prediction was for glutamine synthase (VF_0098), where the model predicted the gene was essential but experimentally it was found to be non-essential. Based on this discrepancy, the modified SMILEY algorithm predicted the draft model was missing glutamine transporter(s) (Figure 2.2d).

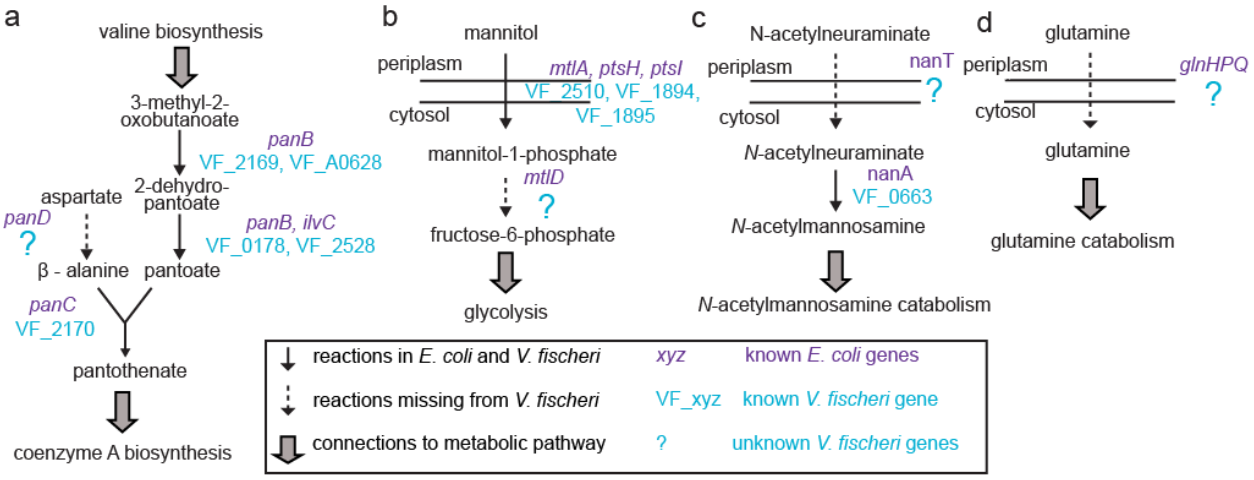


Figure 4.2 Pathways Missing Reactions and Genes in *V. fischeri*. (a) Aspartate 1-decarboxylase (*panD* in *E. coli*), involved in panthothenate and coenzyme A biosynthesis, is missing in *V. fischeri*. (b) Mannitol-1-phosphate 5-dehydrogenase (*mtlD* in *E. coli*), involved in mannitol catabolism, is missing in *V. fischeri.* (c) Transporters for *N*-acetylneuraminate (*nanT* in *E. coli*) and (d) glutamine (*glnHPQ* in *E. coli*) are missing in *V. fischeri*.

To experimentally identify the *V. fischeri* genes encoding the missing aspartate 1-decarboxylase, and mannitol-1-phosphate 5-dehydrogenase, as well as transporters for *N*-acetylneuraminate and glutamine, a specific *E. coli* recipient strain and selective medium were designed for each missing metabolic function (Appendix III, Table AI.S1). The iJO1366 metabolic model was used to demonstrate *in silico* that the growth of each *E. coli* recipient strain requires the specified missing metabolic function in selective medium (Figure 4.3a-d). Here, recipient *E. coli* strains were used because their

metabolism is well characterized, and knockout mutant collections are available [45,46]. For clarity, we will refer to the well-characterized *E. coli* genes using their gene symbols and *V. fischeri* genes using their locus tags. The locus tags of the *E. coli* genes are listed in Appendix III Table AII.S1. A 53 Gbp *V. fischeri* genomic library with approximately 12,000-fold genome coverage was created and transformed into the recipient strains: Δ*panD*, Δ*mtlD*, Δ*nanT*, and Δ*glnP*. After growth selection in each recipient strain's selection media, we found that plasmids expressing VF_0892, VF_A0062, and VF_0668 rescued Δ*panD,* Δ*mtlD,* and Δ*glnP,* respectively, and plasmids independently expressing either VF_0924 or VF_1172 rescued Δ*glnP*.

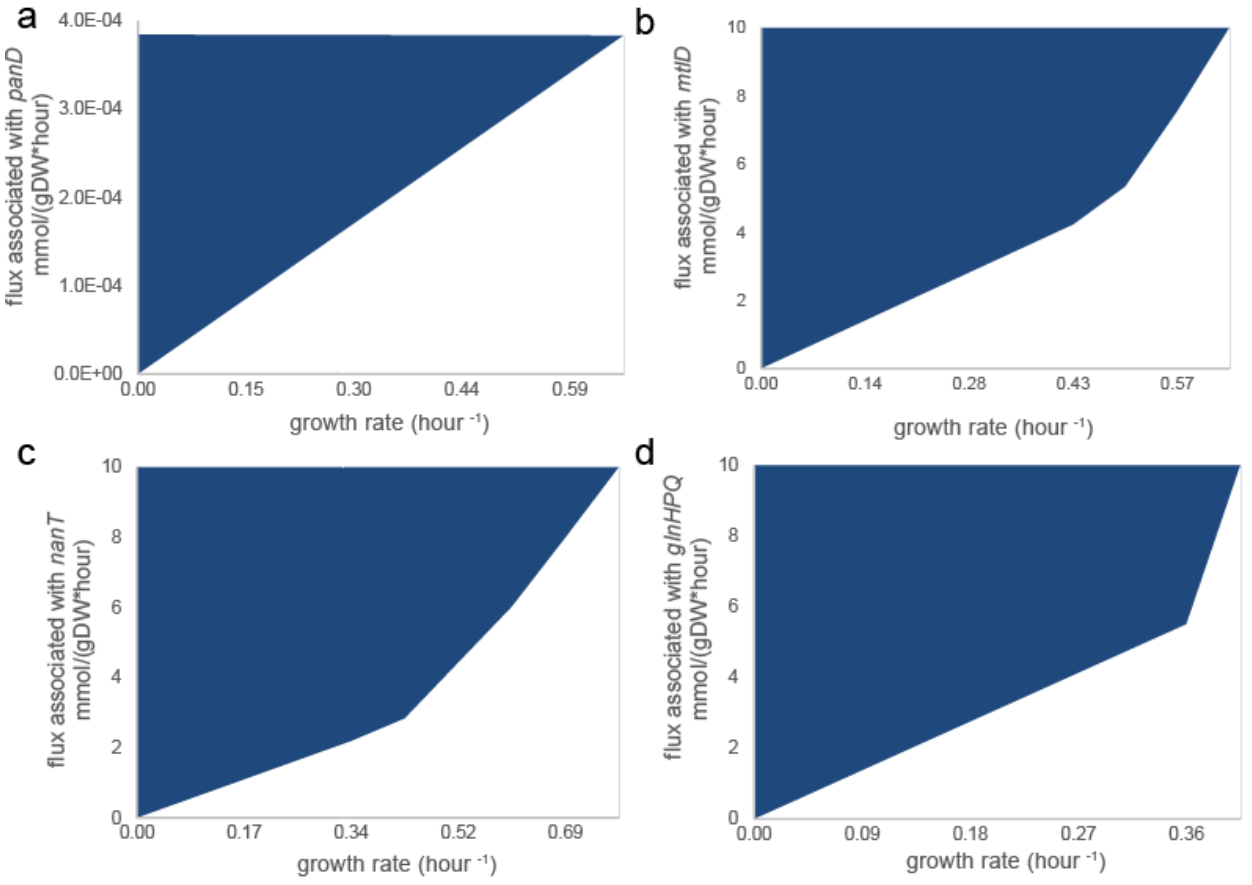


Figure 4.3 Growth Coupling of a Recipient Strain to a Missing Metabolic Function in Selective Medium. Growth dependence for each recipient strain in selective medium calculated using iJO1366 [115]. Feasible combinations of growth rate and a missing metabolic enzyme are shown in blue. In all cases, cell growth is not zero only when there

is flux through the reaction on the y-axis (since no non-trivial solutions exist on the x-axis). The flux limits of oxygen and carbon uptake rates were set at 10 mmol per gram dry weight (gDW) per hour. (a) Aspartate 1-decarboxylase activity (ASP1DC associated with *panD* in iJO1366) is coupled to growth of a Δ*panD* mutant in glucose minimal medium. (b) Mannitol-1-phosphate 5-dehydrogenase activity (M1PD associated with *mtlD* in iJO1366) is coupled to growth of a Δ*mtlD* mutant in mannitol minimal medium. (c) *N*-acetylneuraminate transport (ACNAMt2pp associated with *nanT* in iJO1366) is coupled to growth of a Δ*nanT* mutant in *N*-acetylneuraminate minimal medium. (d) Glutamine transport (GLNabcpp associated with *glnHPQ* in iJO1366) is coupled to growth in a double knockout Δ*glnP*Δ*ansB* in glutamine minimal medium. The *ansB* was not deleted experimentally because it has a low activity with glutamine [117], and Δ*glnP* mutants [118,119] were previously shown to be unable to grow on glutamine.

We individually cloned VF_0892, VF_A0062, VF_0668, VF_0924, and VF_1172 into an empty vector because plasmids in the genomic library can contain fragments that encode more than one gene. All transformed single-gene plasmids enabled growth of recipient strains in the corresponding selective medium (Figure 4.4a-d). This result demonstrated that these genes complemented recipient-strain growth and, thus, are functionally equivalent to the analogous *E. coli* genes. In the experiments with the glutamine transporter, glutamine was supplied as the sole carbon source. Wild-type *E. coli* grows poorly on glutamine as a sole carbon source due to low uptake rates [120]. Overexpression of either VF_0924 or VF_1172 in the *E. coli* Δ*glnP* resulted in a faster growth rate compared to the wild-type *E. coli* strain expressing an empty-vector control (Figure 4.4d).

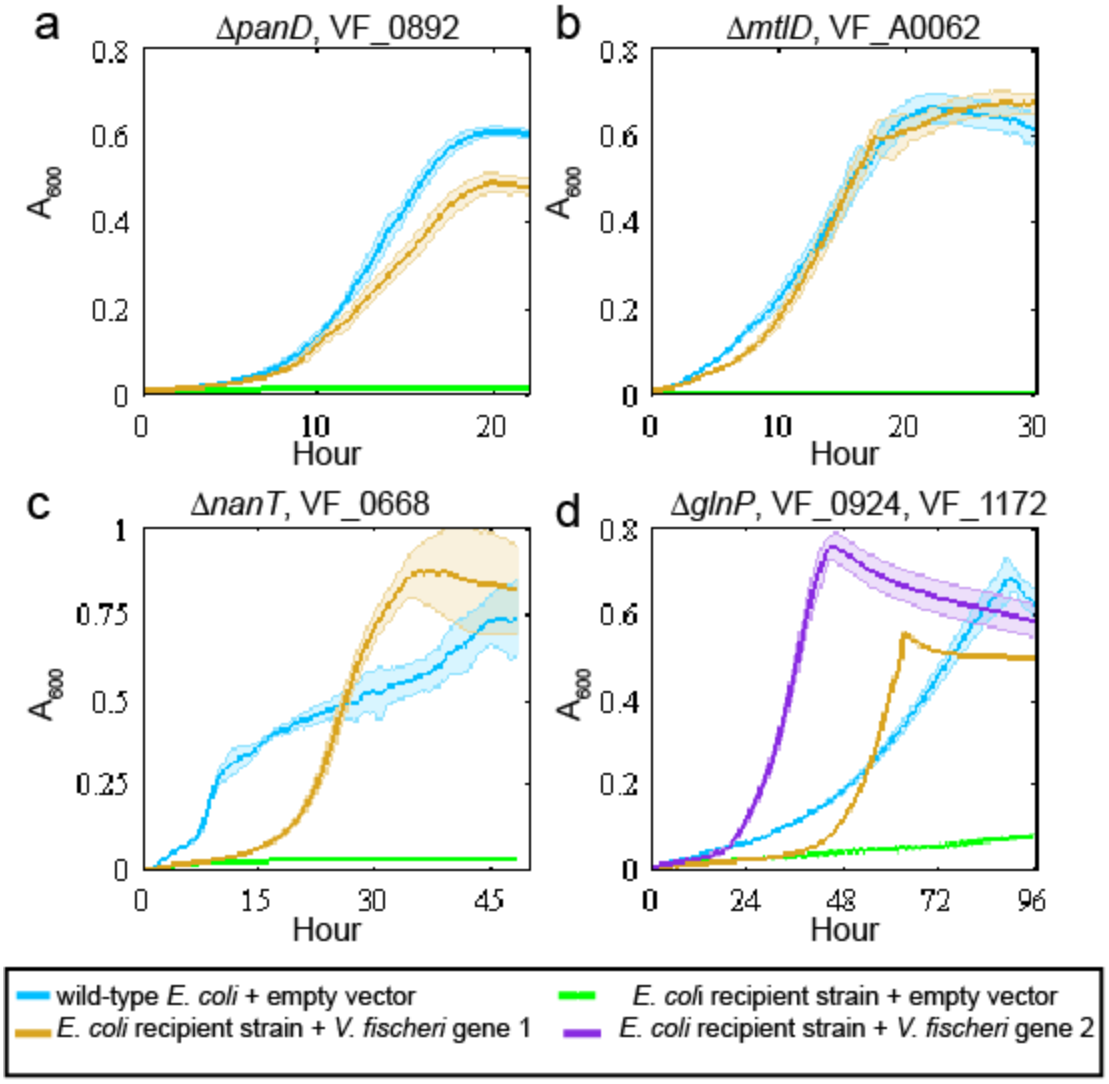


Figure 4.4 Plasmids Expressing the MEGS-discovered *V. fischeri* Genes Enabled Recipient Strains' Growth. Each panel shows the growth curves for the recipient strain listed at the top of each panel. Each solid line represents OD_{600} over time, and the associated colored transparent shade indicates the range of standard deviation across biological replicates (n = 3). The selective medium used for each panel was a MOPS minimal medium supplemented with (a) 20 mM glucose, (b) 20 mM mannitol, (c) 20 mM *N*-acetylneuraminate, and (d) 20 mM glutamine and vitamin supplements (0.05 mM thiamine, 0.05mM niacinamide, and 20 nM biotin). Blue lines represent growth of the parent BW25113 with an empty pZE21MCS vector. Green lines represent growth of the listed recipient strains (Δ*panD*, Δ*mtlD,* Δ*nanT,* or Δ*glnP)* with an empty pZE21MCS vector. Yellow and purple lines represent growth of recipient strains that contain pZE21MCS plasmids expressing the listed *V. fischeri* gene. In (d), the yellow line represents Δ*glnP* + pZEVF0924 strain and the purple line represents Δ*glnP* + pZEVF1172.

The ΔVF_0892, ΔVF_A0062, or ΔVF_0668 *V. fischeri* knockout mutants did not grow in minimal medium (similarly to the selective medium, where DMM instead of MOPS minimal medium was used); however, the *V. fischeri* knockout mutants were complemented by plasmids expressing wild-type copies of VF_0892, VF_A0062, or VF_0668, respectively (Figure 4.5a-c). In addition, the ΔVF_0892 mutant could grow in glucose minimal media if supplemented with pantothenate and/or β-alanine. Better growth of this mutant was observed with addition of 10 mM pantothenate compared to 10 mM β-alanine, possibly due to a slower β-alanine uptake. However, the transporter(s) in *V. fischeri* for β-alanine and pantothenate are, as yet, unknown. The ΔVF_0924ΔVF_1172 double mutant still grew in DMM supplemented with glutamine as the sole carbon source, indicating the existence of other *V. fischeri* glutamine transporters in the genome. We tested whether VF_1172 (annotated as a tyrosine-specific transporter) can also transport leucine, by evaluating growth of *E. coli* BW25113 with a plasmid over-expressing VF_1172 in minimal medium supplemented with glucose and leucine. Over-expression of VF_1172 slowed growth in media supplemented with leucine, a phenotype that can be attributed to leucine toxicity that was previously observed in an *E. coli* K-12 strain over-expressing branched-chain amino acid transporters [121]. Like VF_1172, other amino acid transporters of *V. fischeri* might have broad substrate-specificity.

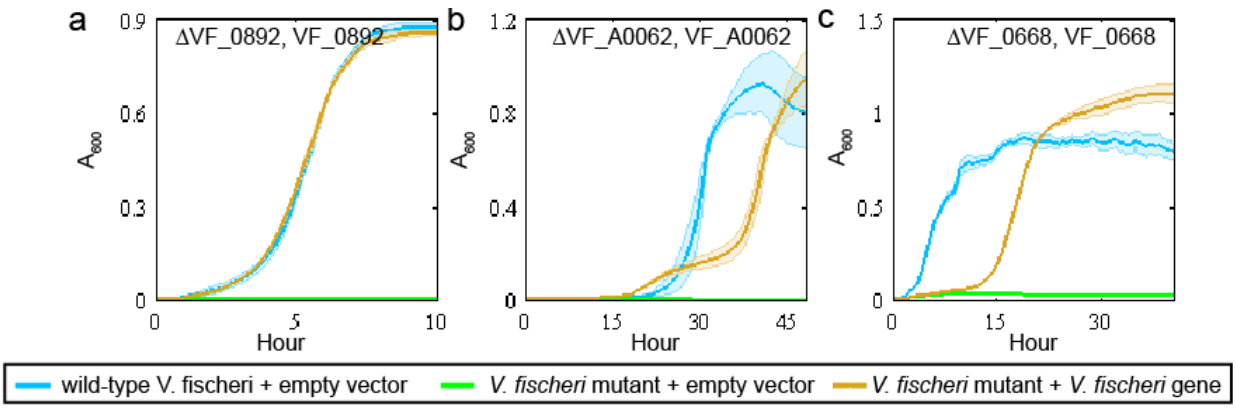


Figure 4.5 Plasmids Expressing *V. fischeri* Genes Complemented Growth of *V. fischeri* Knockout Mutants. Each panel shows the growth curves for a *V. fischeri* mutant complemented with an empty vector or vector expressing the deleted gene. The gene listed at the top of each panel indicates the gene deleted in the mutant and expressed in the complementation plasmid. Each solid line represents OD_{600} over time, and the associated colored transparent shade indicates the range of standard deviation across biological replicates (n = 3). The medium used in each panel was the DMM minimal medium supplemented with (a) 20 mM glucose, (b) 20 mM mannitol, or (c) 20 mM *N*-acetylneuraminate. Blue lines represent growth of *V. fischeri* ES114 with an empty pVSV105 vector. Green lines represent growth of the *V. fischeri* knockout mutants (ΔVF_0892, ΔVF_A0062, ΔVF_0668) with an empty pVSV105 vector. Yellow lines represent growth of *V. fischeri* knockout mutants carrying a pVSV105 plasmid expressing the corresponding *V. fischeri* gene (e.g., ΔVF_0892 + pVF0892).

To further confirm the enzymatic function of VF_0892, we purified a His_6-tagged VF_0892 protein and tested its functionality as an aspartate 1-decarboxylase *in vitro*. Only the reaction condition that contained both aspartate and the VF_0892 enzyme was able to produce β-alanine from aspartate (12.7 ± 0.8 nmol, where reported error is the standard deviation across 3 biological replicates). In contrast, conditions containing aspartate alone, VF_0892 enzyme alone, aspartate and heat-inactivated VF_0892 enzyme, or aspartate and proteins purified from cells containing the empty vector, did not produce any detectable β-alanine (less than 0.125 nmol). Proteins purified from cells containing the empty vector were used as a control to provide information of contaminant proteins. These *in vitro* enzyme assays further confirmed that VF_0892 (1644 bps) can catalyze the conversion of aspartate to β-alanine and, thus, is functionally equivalent to

panD (381 bps) despite their sequence dissimilarity (i.e., BLAST found no significant similarity between the two proteins).

VF_0892 is currently annotated in NCBI as a glutamate decarboxylase (E.C. 4.1.1.15). To compare the activity and substrate specificity of VF_0892 and *E. coli* PanD towards the two potential substrates (aspartate and glutamate), decarboxylase activities were evaluated using the DC-PEPC-MDH-linked assays at pH 8.05. Here, the kinetic parameters were determined from 3 technical replicates and are reported as an average ± SE. The k_{cat} and K_m of the VF_0892 enzyme were 0.075 ± 0.04 µM CO_2/ µM enzyme-sec and 1.44 ± 0.35 mM, respectively, at 28°C, and 0.008 ± 0.001 µM CO_2/ µM enzyme-sec and 1.70 ± 0.56 mM, respectively, at 37°C (Figure 4.6a). The VF_0892 enzyme was around 10-fold more active at 28°C compared to 37°C. Precipitation of purified VF_0892 enzyme was observed over time at 37°C. Greater activity at 28°C is consistent with the optimal growth of *V. fischeri* at 28°C and its intolerance to higher temperatures. *E. coli* PanD showed a k_{cat} of 0.008 ± 0.001 µM CO_2/ µM enzyme-sec and K_m of 1.44 ± 0.33 mM at 37°C (Figure 4.6b). However, it is possible that not all of the purified PanD was post-translationally processed into its active form, resulting in a low k_{cat}. Both PanD and the VF_0892 enzyme showed a much higher reaction rate when aspartate, rather than glutamate, was used as the substrate (PanD: 10-fold; VF_0892: 5-fold at 37°C, and 26-fold at 28°C, Figure 4.6c).

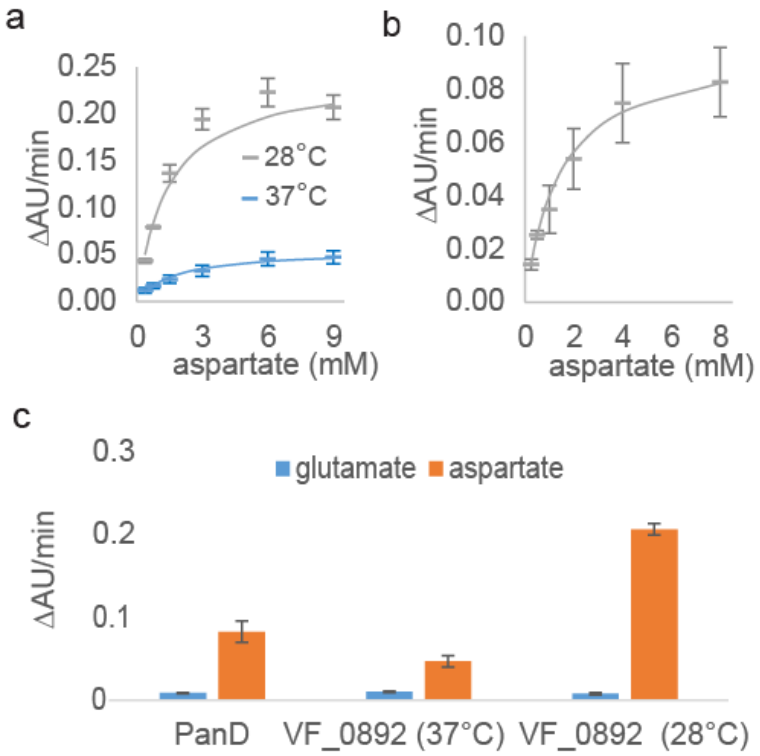


Figure 4.6 Kinetic Characterization and Substrate Specificity of the *E. coli* PanD and the VF_0892 (PanP) Enzymes. Average reaction rates (ΔAU/min) are shown for three technical replicates of the DC-PEPC-MDH-linked assays performed at the specified aspartate or glutamate concentration with standard deviation shown as the error bars. (a) Reactions rates with the VF_0892 enzyme at various concentrations of aspartate at either 28°C (gray) or 37°C (blue). Solid lines show the nonlinear fitting to the Michaelis-Menten equation using KaleidaGraph. (b) Reaction rates of PanD at various concentrations of aspartate at 37°C (c) Reaction rates of PanD with a concentration of 8 mM of either glutamate or aspartate, and reaction rates of the VF_0892 enzyme with a concentration of 9 mM of either glutamate or aspartate at either 28°C or 37°C.

In addition to VF_0892, another *V. fischeri* gene (VF_1064) is currently annotated as a glutamate decarboxylase. Based on BLASTP, VF_1064 has 21% identity to VF_0892 with an E-value of 0.38, but 58% identity and an E-value of 0.0 to the *E. coli* glutamate decarboxylase *gadB*. We evaluated experimentally whether VF_1064 decarboxylates glutamate and/or aspartate. In the DC-PEPC-MDH-linked assays, the purified VF_1064 enzyme showed about a 20-fold higher reaction rate with 10 mM glutamate than with 10 mM aspartate (Appendix III, Figure AII.S2a). With glutamate at pH

8.05 and 37°C, VF_1064 exhibited a k_{cat} of 0.055 ± 0.037 µM CO_2/ µM enzyme-sec and K_m of 61.7 ± 46.4 mM (Appendix III, Figure AII.S2b). Due to the solubility of glutamate, we were not able to test the enzyme kinetics of VF_1064 at glutamate concentrations at or greater than its K_m, resulting in large SEs for k_{cat} and K_m. A pH of 8.05 was used to keep CO_2 produced by the decarboxylase primarily as HCO_3^-, but this pH might not be optimal for VF_1064 because *E. coli* GadB is most active at pH 3.8 [122]. Additional experimental evidence suggests that VF_1064 cannot function as an aspartate 1-decarboxylase; specifically, only VF_0892, and not VF_1064, is essential in LBS medium [54]. Also, the single-gene plasmid expressing VF_1064 did not complement the Δ*panD E. coli* mutant in the selective medium, while the plasmid expressing VF_0892 did (Figure 4.4a).

Based on the measured substrate preferences for the VF_0892 and VF_1064 enzymes, we conclude that VF_0892 is associated with the aspartate 1-decarboxylase reaction, while VF_1064 is only associated with the glutamate decarboxylase reaction in the *V. fischeri* model.

We further tested whether *V. fischeri* knockout mutants (ΔVF_0892, ΔVF_A0062, ΔVF_0668, and ΔVF_0924ΔVF_1172) could successfully colonize the squid light organ in competition with a wild-type strain (we excluded VF_0892 due to its inability to grow in LBS medium without supplementation of β-alanine or pantothenate). We observed no significant colonization phenotypes in our knockout mutants during the initial stage of colonization (24 hours after inoculation). However, VF_A0062 and VF_0668 might play a role in persistence after colonization because transposon insertions in either VF_A0062

or VF_0668 failed to persist in the squid when competing with a pool containing other transposon library mutants and the wild type (after 48 hours of colonization)[54].

4.1.3 Finding Orphan Enzymes by Extending MEGS Strategies

In addition to locating sequence(s) that are responsible for gap-filled reactions or resolve inconsistencies between model predictions and experiment results, we imagine that MEGS can be extended to find sequences of orphan enzymes. Here we propose two strategies for the discoveries of two orphan enzymes, guanosine deaminase (EC 3.5.4.15) and oxaloglycolate reductase (EC 1.1.1.92), that have enzyme activities detected in *Pseudomonas putida* [123]. Following the protocol described in MEGS [58], we constructed a genomic library of *P. putida* with a titer of 370, 000 and average insert size of 1.6 kbp.

Similar to the MEGS approach described in **Chapter 4.1.1** and **4.1.2**, a host strain where the orphan enzyme of interest is essential to growth has to be designed. The first orphan enzyme, guanosine deaminase produces ammonium and xanthosine from guanosine. Because guanosine, ammonium, and xanthosine all are nitrogen sources to *E. coli*, we could knockout some genes to disable *E. coli*'s ability to use guanosine as a nitrogen source such that guanosine deaminase has to convert guanosine to ammonium and xanthosine when only guanosine is supplied as a nitrogen source. The growth-coupled knockout strategies (Δ*xapA deoD ppnP gsk)*—which are free of guanosine phosphorylase (EC 2.4.2.1) and inosine/guanosine kinase (EC 2.7.1.73)—are confirmed with the *E. coli i*JO1366 model to be growth coupled to guanosine deaminase under this specified medium condition but this is yet to be validated experimentally.

In the case of the second orphan enzyme, oxaloglycolate reductase. We designed a host strain growth-coupled to oxaloglycolate, *E. coli* Δ*ttdAB*, free of L-tartrate dehydratase (EC 4.2.1.32) activity. Therefore, when supplying L-tartrate as a sole carbon source to Δ*ttdAB*, L-tartrate can be first converted to oxaloglycoate by L-tartrate dehydrogenase (DmlA), but then oxaloglycolate reductase is required to convert oxaloglycolate to D-glycerate for supporting cell growth. Again, this growth-coupled selection strategy is only verified computationally and needs further experimental validation.

After the growth-coupled *E. coli* Δ*xapA deoD ppnP gsk* and Δ*ttdAB* strains are constructed and validated, we will be able to transform the *P. putida* library into these two strains and select for guanosine deaminase and oxaloglycolate reductase in minimal medium with guanosine as a sole nitrogen source and L-tartrate as a sole carbon source respectively.

4.2 Conclusions

In this work, we developed MEGS to improve gene annotations by combining metabolic modeling with genomic functional selection. Computational models, built from an existing genome annotation, were used to identify missing or incorrect annotations in the current genome annotation, and to design selections (using other organisms) for genes responsible for these missing functions. We successfully identified 5 genes responsible for 4 metabolic functions that were missing from our draft *V. fischeri* metabolic model. Using MEGS, we provided the first experimental evidence that (*i*) VF_0892 functions as an aspartate 1-decarboxylase, (*ii*) VF_A0062 as a mannitol-1-phosphate 5-dehydrogenase, (*iii*) VF_0668 as a *N*-acetylneuraminate transporter, and (*iv*) VF_0924

and VF_1172 as glutamine transporters; importantly, none of these genes are orthologous to the *E. coli* genes with the same functions. These discoveries improved the quality of the *V. fischeri* genome-scale metabolic model, *i*VF846, which has been used in studying *V. fischeri* metabolism and its symbiotic relationship in the squid light organ, and especially during the habitat transition between seawater and the symbiotic niche [53]. Additionally, we have extended MEGS and designed strategies that are ready to be implemented experimentally for discovering two orphan enzymes, guanosine deaminase and oxaloglycolate reductase.

MEGS leverages both computational and experimental techniques to provide functional annotations for genes encoding enzymes and transporters. Metabolic genes are responsible for the physiological and biochemical states of a cell, and knowledge of their functions is critical for understanding and controlling cell behavior. By taking advantage of metabolic modeling, MEGS identifies errors and omissions in existing genome annotations due to either a lack of experimental evidence or prior knowledge in databases, and designs experiments to correct these errors. Because MEGS uses genomic functional selections to find genes instead of sequence similarity or genomic context, it can discover genes with unique sequences and/or genes that have not been studied in the laboratory. MEGS is experimentally and computationally inexpensive and efficient. A draft genome-scale metabolic model can be prepared automatically using available software platforms in only a few hours [108,124–126]. Metabolic reactions missing from such a draft model or associated with the wrong genes, as well as genomic functional-selection strategies, can also be identified within a few hours [14]. The genomic library used in our method (which takes 2 days to construct) contains information from

across the entire genome, so that all the genes in the library go through the selection simultaneously. Once a library is created, MEGS cycles can be repeated to search for additional genes associated with different missing reactions. Similarly, once a recipient strain is constructed, it can be used to select for genes responsible for a particular metabolic reaction from multiple genomes (using separate genomic libraries). The time required to build the recipient strains depends on the number of gene additions and deletions needed; however this time can be significantly reduced by using existing mutant strain collections [44,46,47]. Functional selection from the genomic library via growth complementation of the recipient strain in a selective medium is fast (1 to 3 days) and also provides direct experimental evidence of gene functions. One cycle of the entire MEGS process can be completed within a week using existing recipient strain collections.

MEGS offers an alternative and complementary approach to bioinformatics based methods for predicting gene functions. In retrospect, some of the genes we identified using MEGS are also top-scoring candidates that have been or could be derived bioinformatically using various genomic context-based methods. VF_0892 was already tentatively suggested to be a member of the pyridoxal-dependent aspartate 1-decarboxylases (TIGR03799) by partial phylogenetic profiling (29). This protein family was given a suggested name of PanP, to distinguish it from a non-orthologous family of aspartate 1-decarboxylase (PanD TIGR00223), which is pyruvoyl-dependent and more widely distributed than PanP. PanP is present in a number of marine bacteria; however, no direct experimental evidence was available previously to support its annotation as an aspartate 1-decarboxylase. We found that PanP was not properly incorporated in six manually curated genome-scale metabolic models. Five of the models included an

aspartate 1-decarboxylase reaction without any associated genes [127,128], and one associated PanP with L-cysteate, 3-sulfino-L-alanine, glutamate, and aspartate decarboxylase reactions [57]. Similarly, VF_A0062 could have been predicted as a mannitol-1-phosphate 5-dehydrogenase candidate using bioinformatics methods because it is located in the same operon as *yggD* (VF_A0063), and YggD has been characterized as a mannitol operon repressor in *Shigella flexneri* (34). VF_0668 is a predicted member of SiaR regulon (controlling sialic acid degradation) according to the RegPrecise database [64], and it was tentatively annotated as a possible sialic acid transporter (permease), NanT. The SiaR regulon in *Haemophilus influenza* includes a different tripartite ATP-independent periplasmic transporter [129]. However, the function of the *V. fischeri* NanT has not yet been experimentally characterized.

Similar to operon- and regulon-based bioinformatics methods, MEGS does not depend on sequence similarity to well-characterized genes. Annotations based solely on sequence similarity may lack detailed functions when the unknown sequence is not similar to a characterized gene (e.g., conserved hypothetical proteins) or may be incorrect if the gene is similar to a gene that is incorrectly annotated in sequence databases. In the case of VF_A0062, we demonstrated that this gene actually encodes a mannitol-1-phosphate 5-dehydrogenase even though it shares high sequence similarity with other annotated L-sorbose 1-phosphate reductase genes.

MEGS and bioinformatics-based approaches have different strengths and limitations and, as a result, can complement each other. For example, MEGS can complement bioinformatics-based approaches when top-scoring candidates do not exist. Sometimes correlations between genes with unknown function and genes with known

functions may not exist, and such correlations can still lead to erroneous and/or non-specific function assignments using bioinformatics methods alone (8). Another strength of MEGS compared with the bioinformatics-based approach is that genes identified from MEGS already have direct experimental evidence for their functions from the genomic-library selection experiments, while bioinformatics-derived gene functions must be tested in subsequent experiments. MEGS can be applied to organisms for which there are currently no genetic tools, opening up ways to evaluate their gene functions experimentally in another host. Some limitations of MEGS include that (i) it requires heterologous expression in the recipient strain, which might not be optimal, (*ii*) the recipient strains might be difficult to construct (*e.g.*, essential genes cannot be deleted unless growth can be complemented by nutritional supplementation), and (*iii*) if multiple genes are responsible for a missing metabolic function, they must be co-localized on the chromosome. Additionally, MEGS might identify genes with low promiscuous activities that, when over-expressed, can complement growth defects associated with essential metabolic functions. Recent approaches for improving heterologous gene expression [130] and conditional mutation systems [131] are likely to help overcome some of these limitations. Ultimately, more detailed biochemical characterization of enzymes identified using bioinformatics or MEGS might be needed to confirm the physiological functions of gene products. However, these computational and experimental approaches are useful for identifying which genes to evaluate biochemically.

As more genome sequences become available, we speculate that MEGS will be successfully applied to further discover the roles of uncharacterized genes and improve our understanding of metabolism in a variety of both familiar and uncharacterized

microorganisms. Newly discovered gene functions will propagate through genome databases as they are used by other existing approaches to improve annotations of genes in additional organisms.

Chapter 5

Model-Guided Metabolic Discoveries in *Zymomonas mobilis*

Zymomonas mobilis is a model system for production of biofuels and biochemicals [132,133]. As a natural ethanologen, *Z. mobilis* has a strong resilience to ethanol and many lignotoxins present in renewable feedstocks, extremely fast glucose uptake rates, and relatively low biomass production due to its special metabolism capabilities. Instead of using the Embden-Meyerhof-Parnas (EMP) pathway for glycolysis, it uses the Entner-Doudoroff (ED) pathway. Our group has started the investigation of the metabolism of the industrial *Z. mobilis* strain ZM4 (ATCC 31821) via the approaches of genome-scale metabolic modeling and metabolic flux analysis to facilitate better metabolic engineering for biofuel production. A previous student of our group, Wai Kit Ong, drafted a genome-scale metabolic model for *Z. mobilis* ZM4, which located multiple gaps in our knowledge about this organism. In this project, we provided new insights about *Z. mobilis* ZM4 metabolism after validating hypothesized reasons behind some of these gaps using growth experiments and MEGS [58], my previously published approach for gene search.

5.1 Results and Discussion

5.1.1 Discovering Genes for Reactions without Gene Associations

In the draft ZM4 genome-scale metabolic model, there are almost 20% reactions (146 out of 747 total reactions) without gene associations, i.e., non-GPR reactions. Except exchange and sink reactions which are non-GPR reactions required for establishing mass balance in the model, non-GPR reactions with a specific metabolic function may indicate knowledge gaps where gene sequences for these reactions are

missing. Using the model-enabled gene search (MEGS) approach that we previously developed [58], we were able to discover gene sequences responsible for three non-GPR reactions (chorismate-pyruvate lyase encoded by *ubiC* in *E. coli*, erythronate-4-phosphate dehydrogenase, encoded by *pdxB* in *E. coli*, and pimeloyl-ACP methyl ester esterase encoded by *bioH* in *E. coli*) that were gap-filled into the draft model during the curation process (Table 5.1). Following the MEGS protocol, first an *E. coli* host strain and selective medium pair for each reaction was designed such that the reaction is essential to the host strain in selective medium. Then a genomic library of *Z. mobilis* was created and transformed into the host strain also following the MEGS protocol. The average insert size of the library was ~1.4 kbp and the titer of the library was 2.1×10^5 colony forming units (CFUs). Finally, plasmids contained in individual colonies that survived the selective medium were sequenced. *E. coli* knockout mutants Δ*ubiC*, Δ*pdxB*, and Δ*bioH* were used as host strains, and the selective media used during the selection for these genes were MOPS minimal medium with 20 mM malate, MOPS minimal medium with 20 mM glucose, and M9 minimal medium with 20 mM glucose respectively. Plasmids that contained a single gene of ZMO1008 or ZMO1916 complemented the growth of Δ*pdxB* and Δ*bioH* respectively. Therefore, ZMO1008 is likely to encode an erythronate-4-phosphate dehydrogenase and ZMO1916 is likely to encode a pimeloyl-ACP methyl ester esterase. In the case of chorismate-pyruvate lyase, both ZMO0562 and ZMO0563 were on plasmids that complemented the growth of Δ*ubiC* in the selective medium. ZMO0563 was likely the true chorismate-pyruvate lyase because when ZMO0562 and ZMO0563 were cloned separately into two different plasmids only ZMO0563 was able to complement the growth of the mutant. To our knowledge, we were the first to discover the *Z. mobilis*

chorismate-pyruvate lyase and erythronate-4-phosphate dehydrogenase. Our work also provided direct evidence for the *Z. mobilis* pimeloyl-ACP methyl ester esterase, which has been previously inferred from pooled mutant fitness data, where the ZMO1916 mutant showed high cofitness scores with mutants of biotin synthase (ZMO0094; r = 0.95) and dethiobiotin synthase (ZMO0095, r = 0.8) [98]. Although previous mutant fitness data supported that ZMO1916 is involved in biotin synthesis, unlike our experiments, these data did not provide direct evidence of the exact function of ZMO1916.

Experimental Activity and Pathway	Zymomonas Gene	KEGG Annotation
UbiC (Chorismate-pyruvate lyase); Ubiquinone synthesis	ZMO0563	Chorismate mutase
PdxB (Erythronate-4-phosphate dehydrogenase); Pyridoxal 5'-phosphate synthesis	ZMO1008	FAD linked oxidase domain protein
BioH (Pimeloyl-ACP methyl ester esterase); Biotin biosynthesis	ZMO1916	Conserved hypothetical protein

Table 5.1 Genes Found by MEGS for Non-GPR Reactions. KEGG annotation is listed here as a reference because the draft model was reconstructed from the KEGG annotation of *Z. mobilis* ZM4 genome.

5.1.2 Investigating "Missing" Enzymes in *Z. mobilis* Tricarboxylic Acid Cycle

In the tricarboxylic acid (TCA) cycle of our *Z. mobilis* model, there lack homologs to most subunits of the *E. coli* fumarate reductase (*frd*) and succinate dehydrogenase (*sdh*)—except ZMO0569, an annotated succinate dehydrogenase cytochrome b subunit—that commonly exist in many other bacteria to lead to the production of succinate and eventually the essential metabolite succinyl-CoA for amino acid biosynthesis. The Frd/Sdh reactions were included in two published ZM4 genome-scale metabolic models [134,135]. However, more evidence would be necessary for including Frd/Sdh in a model. Therefore, MEGS was applied to search for genes associated with Frd and Sdh using an

E. coli frd mutant—Δ*frdA*, Δ*frdB,* Δ*frdC*, or Δ*frdD*—as a host strain in a MOPS minimal medium with 20 mM fumarate and 40 mM glycerol anaerobically. However, the *Z. mobilis* genomic library was not able to rescue any of the *frd* mutants. This result indicated that under the tested conditions *Z. mobilis* is unlikely to have Frd/Sdh activities and might instead produce succinate via the biosynthesis of NAD, UTP, and CTP only without using the TCA cycle.

In addition to fumarate reductase and succinate dehydrogenase, we attempted to find an isozyme for another common TCA cycle enzyme, fumarase. There is a discrepancy of model predicted no growth and experimentally predicted growth of the fumarase ZMO1307::Tn5 mutant. This discrepancy was especially interesting to us because it was related to the missing Frd/Sdh and unknown directions of fluxes between fumarate and succinate in the *Z. mobilis* TCA cycle. In *E. coli*, there exist three fumarase isozymes. FumA and FumC are expressed under aerobic conditions while FumB is only expressed under anaerobic conditions. We used *E. coli* Δ*fumAC* as a host strain for selection of potential *Z. mobilis* fumarase isozymes aerobically in fumarate MOPS minimal medium. Two genes that showed up from the genomic library plasmids that independently complemented the *E. coli* Δ*fumAC* were ZMO1307 and ZMO1404. ZMO1404 is a sigma factor in *Z. mobilis*. We suspected that the expression of ZMO1404 turned on the expression of *E. coli fumB* under the aerobic condition. We cloned a ZMO1404-containing plasmid and transformed the plasmid into *E. coli* Δ*fumAC* and *E. coli* Δ*fumACB* mutants. This plasmid successfully complemented the Δ*fumAC* mutant but not the Δ*fumACB* mutant aerobically in fumarate MOPS minimal medium, showing that ZMO1404 indeed is able to turning on the expression of the anaerobic *E. coli fumB* under

an aerobic condition. Finally, we tested the growth of a ZMO1307::Tn5 mutant and detected fumarate secretion via high-performance liquid chromatography (HPLC)—fumarate secretion by wild type was not detected—supporting the hypothesis that ZMO1307 is the only *Z. mobilis* fumarase. This experiment also showed that *Z. mobilis* has a fumarate transporter. When added to the model, the fumarate transporter can resolve the no-growth and growth misprediction by the model.

5.1.3 Investigating Flux Coupling Analysis Modules with Poor Cofitness in Pooled Mutant Assays

A previous Reed group member, Wai Kit Ong applied flux coupling analysis to the draft ZM4 model [13] and calculated—based on network structures—modules of genes that ideally have related or similar metabolic functions and a high average cofitness score in pooled transposon mutant assays [98]. Out of all 36 modules that have more than two genes present in the transposon datasets [98], we analyzed each of the 6 modules with a poor average cofitness score, and in some cases, examined the modules using growth experiments and model-enabled gene search (MEGS) [58].

The first experimentally tested module, M56, includes seven relevant mutants and 11 reactions involved in folate biosynthesis. In M56 not only do most mutants have poor correlation in the fitness scores, some growth phenotypes of the mutants in the pooled mutant assays (under a glucose minimal medium anaerobic condition) do not match with model predictions (model essential and experimental non-essential). Several issues showed up in the laboratory experiments. First, we found that some isolates used in the pooled mutant assays were incorrect. For example, one isolate of the small glutamine amidotransferase subunit of the 4-amino-4-deoxychorismate synthase (ADCS), ZMO0113::Tn5 (isolate E9 in plate 5), was verified as a mixture of the wildtype and the

correct mutant, and another isolate of the ZMO0113::Tn5 (isolate A12 in plate 30) was the correct mutant. Second, medium carry-over could be another reason that caused poor cofitness scores in M59. When we tested the growth of a correct ZMO0114::Tn5 isolate —the mutant for the large catalytic subunit of ADCS—in ZMMG under anaerobic conditions, it showed a weak growth up to OD_{600} 0.2 after 24 hours. However, when we transferred the culture to fresh ZMMG media, there was no growth after 48 hours. Therefore, the model accurately predicted that ZMO0114 is an essential gene and the pooled mutant assays showed an inaccurate result probably due to medium carry-over. Opposite of ZMO0114, ZMO0113 has a true growth phenotype that was wrongly predicted by the model as no growth, indicating a possible isozyme of ZMO0113 might exist. Finally, we used the MEGS approach to search for alternative isozymes of ZMO0113 and ZMO0114. *E. coli* Δ*pabA* and Δ*pabB* strains were used as a host strain for the search of ZMO0113 and ZMO0114 isozymes respectively in aerobic M9 glucose minimal medium. Genomic library plasmids containing either ZMO0113 or ZMO0201 rescued the growth of Δ*pabA* while Δ*pabB* was only rescued by plasmids containing ZMO0114. In the KEGG database [62], ZMO0201 is annotated as the glutamine amidotransferase of anthranilate synthase, and ZMO0201 and ZMO0113 are similar in several ways. First they share the same substrates, chorismate and glutamine. Second, both of them are glutamine amidotransferases. In fact, on NCBI [60], the updated annotation of ZMO0201 is a dual functional aminodeoxychorismate/anthranilate synthase component II, which further supports our MEGS result that ZMO0201 is an isozyme of ZMO0113.

M53, also experimentally tested, includes three mutants and three reactions that convert 3-phosphoglycerate into serine. All three mutants in this module were experimentally essential in glucose minimal medium under anaerobic conditions, but were incorrectly predicted to be nonessential by the model because the metabolic model has additional routes to make serine from glycine through the glycine hydroxymethyltransferase (GHMT2r) reaction. In this case, making the GHMT2r reaction irreversible in the model would accurately capture the essentiality of the serine biosynthesis pathway. Out of the three reactions, phosphoserine phosphatase and D-3-phosphoglycerate dehydrogenase correlate relatively well (cofitness = 0.498), and phosphoserine aminotransferase does not correlate well with any of the two reactions (cofitness scores are 0.089 and 0.292). Using MEGS, we confirmed if phosphoserine aminotransferase is indeed encoded by ZMO1684. *E. coli* Δ*serC* mutant was used as a host strain for selection in MOPS glucose minimal medium. However, ZMO1684 was the only gene on plasmids that complemented the growth of mutant and no other genes were found. We suspect that ZMO1684 does not correlate well with the two other genes in the serine biosynthesis module because it encodes a bifunctional enzyme involved in pyridoxine biosynthesis.

Interestingly, the pyridoxine biosynthesis module that ZMO1684 is involved in, i.e., M44, also has a poor average cofitness score. It is consisted of three genes and five reactions. In M44, ZMO1684 has a cofitness score with ZMO1708 of 0.752, which is much higher than its cofitness scores with genes in M53. Similar to M53, M44 is an example of a module with a poorly correlated mutant bringing the average cofitness down. While the cofitness for two mutants in this module (ZMO1684 and ZMO1708) was 0.75, ZMO1313

has a cofitness around 0.37 with each of these two mutants. ZMO1313, annotated as 4-hydroxythreonine-4-phosphate dehydrogenase, is associated with pyridoxine 5'-phosphate synthase (PDX5PS) reaction. ZMO1313:Tn5 is a NGU (model No Growth and experiment Undetermined) mutant in glucose minimal medium under anaerobic condition. We confirmed one of the only two mutants in the entire BarSeq library was a wildtype (a PCR band corresponding to wildtype copy of the gene indicating the transposons were not in the expected genes). Given the small number of ZMO1313 transposon mutants in the library, the presence of wild type contamination, and conflicting results from two pooled mutant experiments (the experimentally supported ZMO1313 essentiality is in the experiment Undetermined category), we suspect that ZMO1313 is a true essential gene.

We did not experimentally tested the three out of the six modules with a poor cofitness score, but suggested reasons behind the poor scores based on the network structure. One of the three modules, M49 includes 4 relevant genes involved in isoprenoid biosynthesis. In M49, ZMO1851—a flavodoxin—is associated with many other reactions besides the 2C-methyl-D-erythritol 2,4 cyclodiphosphate dehydratase (MECDPDH5) reaction. The ZMO1851::Tn5 mutant has very poor cofitness with the three other mutants in this module, and was the only mutant associated with this module whose growth phenotype was mispredicted by the model as NGG (model No Growth and experiment Growth) in glucose minimal medium anaerobic conditions. A different flavoprotein might be associated with this reaction or the mutant may have both wildtype and disrupted copies of the ZMO1851 gene.

The last two modules with low average cofitness, M48 and M60, include mutants that were predicted and found experimentally to be nonessential under glucose minimal

media conditions. M48 includes three relevant mutants involved in energy and nitrogen metabolism. All three mutants are components of the nitrogenase (NIT1b) reaction and were only essential in nitrogen fixation experiments (2 out of 492 pooled mutant experiments). Module M60 includes three mutants involved in thiamin biosynthesis. The ZMO0172 mutant's fitness was weakly and negatively correlated with the two other mutants in the M60 module (ZMO0738 and ZMO1834). One possible explanation for the low cofitness scores in these two modules is that the reactions were not needed in most conditions, since thiamin and ammonia were present in the media of most transposon experiments analyzed here.

5.2 Conclusions

Unlikely *E. coli* metabolism which has been well-studied for years by scientists, the metabolism of *Z. mobilis*, even key features of *Z. mobilis* central metabolism, are unclear to us. Here, we compared the predictions of a *Z. mobilis* genome-scale metabolic model to experimental observations of mutant fitness and located parts of the metabolic network where knowledge gaps exist. Using our previously developed MEGS approach, we found three genes (ZMO0563, ZMO1006, and ZMO1916) that are responsible for reactions initially without GPR associations in the model. Additionally, we experimentally investigated two "missing" TCA cycle enzymes and several flux coupling modules of poor average cofitness scores—which are associated with folate, serine, and pyridoxine biosynthesis—and discovered an isozyme of the glutamine amidotransferase component in the 4-amino-4-deoxychorismate synthase (encoded by ZMO0201).

Chapter 6

Conclusions

6.1 Future Directions

6.1.1 Standardized Curation and Continued Improvement of Genome-Scale Metabolic Models Using High Quality Genome Annotation

The genome-scale metabolic model for *V. fischeri* ES114, *i*VF846, described in **Chapter 2**, has demonstrated its utility for making novel discoveries in *V. fischeri* metabolism and the impacts of the metabolism on squid and Vibrio interactions. However, this model is not perfect yet, and the following aspects could be very important for improving genome-scale models in the future:

First, a standardized pipeline for generating draft genome-scale model is yet to be designed. Although several software tools have been developed for rapid reconstruction of a draft model (**Chapter 1.2)**, not all these tools are mass- and charge-balanced or have correct reaction directions or GPR associations. Each tool also has its own identifiers for reactions and metabolites, making comparison among models difficult.

The in-house MATLAB scripts that the CS master student, Mohit Hotwani, coded and I designed is different from the software tools mentioned above because our scripts create a new model by copying information from a well-curated model. Its biggest advantage is that if the original model is mass- and charge-balanced and have correct information of reaction directions the new model will contain the same information. It is also a relatively simple tool that can create a new model from a model template in a few minutes. The biggest limitation of our approach is that only one model is supplied as a template for creating the new model. While we could extract high quality information from

the template, we might lose unique metabolic features of the new organism of interest, especially if the template and the organism of interest do not share much common metabolic reactions and pathways. Also the extraction of information from the template model to the new model is limited by the number of homologous gene pairs that exist between the two organisms. Our definition of homologous genes based on KEGG orthology identifiers (**Chapter AI.1.4**) is rather conservative. And gene pairs that do not meet our definition of homologous genes might still have the same or very similar functions, which causes missing reactions and genes in the draft model that can only be added by careful but laborious manual curation.

Second, continued manual curation efforts are needed in the long term. Due to the cost of labor, only a fraction of current genome-scale metabolic models are manually curated to eliminate basic errors in the model while many others are generated entirely with a software tool. Even fewer models, except some for model organisms and organisms of special interest, have been or will be improved continuously with the up-to-date knowledge of the organism after the initial model was created, due to limited interests and incentives in publishing an update of a model. Currently many manually curated models are generated by systems biologists or modelers who can relatively quickly finish a modeling reconstruction procedure but they might not be very familiar with the organism of interest or have little time to improve the model in the long term. Ideally, manual curation of a model should be a long term process because it would be the most valuable for biological discoveries if it is done continuously over the years by a subject-matter expert of a specific organism. Trainings of genome-scale metabolic modeling knowledge

for scientists specialized in a specific organism might help to facilitate continued efforts in model curation.

Finally, improving gene annotation quality is key to improving the quality of genome-scale models. Today, even in the genome of the best studied model organism, e.g. *E. coli*, there is still a significant fraction of genes that are not well-annotated. As seen from **Chapter 2.1.3,** databases and bioinformatics pipelines for gene annotation give results that are sometime similar but sometimes quite different. A comparison and validation of annotation results across multiple databases and pipelines is lacking in existing software tools for draft model reconstruction. There is also a need for novel gene annotation methods that could be either bioinformatics tools, high-throughput experimental tools, or combined experimental and computational tools like the MEGS approach described in **Chapter 4** and **Chapter 5**.

Gene annotation could be a difficult process, which is sometimes caused by genes and reactions that simply have never been studied but at other times caused by newly-characterized genes that have been reported in literature but not added to databases. Because of this, a researcher might need a thorough literature search in addition to a BLAST search to accurately understand the function of a gene with a well-characterized homolog. Collaborative efforts to compile information of newly-characterized genes and proteins would greatly benefit genome-scale metabolic modeling and also laboratory research.

6.1.2 Understanding Transcriptomic Data via Genome-Scale Modeling and Enrichment Analysis

Previously, our group has focused on constraint-based genome-scale modeling approaches to study the metabolism of various organisms. While it is great for

applications such as making novel metabolic discoveries, resolving knowledge gaps, and designing strains for producing a chemical of interest, it might not be the optimal modeling tool for analyzing certain datasets. For example, when transcriptomic data was used for constraining intracellular flux predictions from a genome-scale metabolic model, many algorithms did not outperform a simple flux balance analysis [136]. We have also applied transcriptomic data for making qualitative instead of quantitative predictions, *i.e.*, sugar uptake by *V. fischeri* in the juvenile squid (**Chapter 3.1.1**). Although we were able to show the evidence of two substrates, i.e., *N*-acetylneuraminic acid and glycerophosphodiesters, were used by *V. fischeri* in the juvenile squid, we found that calculated flux coupling modules are highly dependent on medium conditions—which are often not well known or hard to be simulated *in silico*—and very few differentially expressed genes in transcriptomic datasets might overlap with genes in flux coupling modules simulating a rich medium condition, showing the limited utility of this approach in analyzing transcriptomic data.

As demonstrated in **Chapter 3.1.2**, the popular enrichment analysis, is a better tool for analyzing transcriptomic data than genome-scale modeling approaches. However, genome-scale modeling can still use hypothesis generated from an enrichment analysis and be used for further *in silico* experimentation (**Chapter 3.1.3**).

We have also noticed that enrichment analysis approaches are not perfect either. For example, results derived from these methods are dependent on the genes defined in the gene sets, which are different in different databases. For example, KEGG [62] combines amino acid biosynthesis and catabolism genes together in a single gene set, and therefore an enrichment analysis with KEGG might not show a differentially

expressed amino acid biosynthesis if only biosynthesis but not catabolism genes are differentially expressed in the datasets. In **Chapter 3.1.2,** we chose the BioCyc gene sets [61], which are better curated than KEGG gene sets, for our enrichment analysis. Also, the results of enrichment analysis are dependent on the type of data input, i.e. log2 fold changes (FCs) or p-values for the log2 FCs. In **Chapter 3.1.2**, we chose a popular enrichment analysis method, gene set enrichment analysis (GSEA) [91], which ranks all gene sets and shows up- and down-regulation of each gene set based on log2 FCs.

6.1.3 Gene Search and Enzyme Engineering in Growth-Coupled Host Strains

A great utility of genome-scale metabolic modeling is designing growth-coupled strains for metabolic engineering by using forced-coupling algorithms (**Chapter 1.1.2)**. MEGS combines forced-coupling algorithms (**Chapter 1.1.2**), gap-filling (**Chapter 1.2**) and genomic library selection and allows the search of gene sequence(s) responsible for a reaction of interest that is essential to the growth of a host strain. To our knowledge, a combined genome-scale modeling and experimental approach had not been applied for gene search applications before the publication of our MEGS approach (**Chapter 4.1.1**), which discovered *V. fischeri* sequence(s) that might not be found easily using a traditional, bioinformatics gene annotation tool such as BLAST (**Chapter 4.1.2**). We demonstrated in *Z. mobilis* that MEGS approach can also lead to novel metabolic discoveries (**Chapter 5**). Additionally, we have designed strategies that extend MEGS for searching orphan enzymes (**Chapter 4.1.3**).

Similarly, growth-coupling can be applied in enzyme engineering, although tricky to implement sometimes. For example, we have tried to engineer an *E. coli* novel 1-deoxy-D-xylulose-5-phosphate synthase (nDxs)—RibB(G108S)—with higher activities by

selecting error prone PCR library mutants in a host strain of Δ*dxs* growth coupled to nDxs. Because Δ*dxs* is not viable without nDxs, a plasmid containing nDxs mutants of higher activities would complement Δ*dxs* to grow faster than mutants with lower activities. And we chose a plasmid backbone with a low copy number and strict promoter so that we could limit the nDxs concentration to a level that the growth of Δ*dxs* with a poor nDxs was very slow, and ideally Δ*dxs* that grew faster should have obtained nDxs mutants with higher activities from the error prone PCR library. However, in reality, random mutations on the plasmid backbone or even the host strain genome occurred to allow a fast growth of Δ*dxs*. Unfortunately, such random mutations occurred very frequently in our experiments likely due to the low copy number, strict promoter, and very low inducer concentration that were used to control nDxs concentration, making the discoveries of nDxs mutants that truly have higher enzyme activities difficult.

For both gene search and enzyme engineering using growth-coupled host strains, multiple challenges exist. One such challenge is that sometimes generating a growth-coupled strain is time consuming or simply impossible. Current forced-coupling algorithms that are mixed-integer programming problems are computationally intensive. Testing different medium conditions for a growth-coupling design might take more than a few hours. Therefore, faster forced-coupling algorithms have to be designed to facilitate gene search and enzyme engineering in growth-coupled host strains. Additionally, such design might require multiple gene knockouts, some of which might be lethal when implemented experimentally. In other cases, unknown reactions and pathways exist, and the gene knockouts designed computationally might not be sufficient to create a growth-coupled phenotype.

6.2 Concluding Thoughts

Hopefully, via this thesis, I have convinced you of the great utility of genome-scale metabolic modeling combined with experimental methods in discovering novel metabolic capabilities of an organism of interest. As I am writing this thesis that summarizes and concludes all my Ph.D. work, I am also imagining that the future of genome-scale metabolic modeling might look very different from what I have done here:

There might exist a super data center that stores, validates, and consolidates all genome-scale metabolic models, and communicates with multiple servers with gene annotation pipelines to allow a comparison of annotation results for reconstructing a genome-scale metabolic model. A researcher studying a specific organism will be trained on this data center to download a model or some specific information that he/she needs, and at the same time record novel discoveries from his/her experiments in this data center for sharing with other researchers. In the future, as the advancement of data acquisition and synthetic biology technologies, many more types of datasets will be incorporated in to genome-scale metabolic modeling algorithms and model-predicted results will be rapidly validated experimentally. Genome-scale metabolic modeling might also be part of a multi-dimensional modeling approach, which will interface with many other modeling approaches, including the popular statistical modeling, metabolic flux analysis, and machine learning methods.

*i*VF846 Supplementary Materials

AI.1 Methods

AI.1.1 Bacterial strain and growth conditions

Unless otherwise specified, wild-type *V. fischeri* ES114 was grown at 28°C with shaking at 225 rpm in 12 by 75 mm culture tubes containing 3ml of either Luria-Bertani salt (LBS) medium [137] or defined minimal medium (DMM) [58] with 20 mM glucose as the sole carbon source.

AI.1.2 Cell Dry Weight and Biomass Composition Measurement

Usually cell dry weight is measured by washing cells with deionized (DI) water to get rid of medium components (including salts and sugars) and then weighing the cells that retained on a filter. However, because *V. fischeri* cells lyse in DI water, we developed an evaporation procedure for the accurate measurement of *V. fischeri* cell dry weight. First, about 50 mL of *V. fischeri* culture grown to the exponential phase in DMM was washed twice with fresh medium containing no carbon source. Then the washed cell pellet was suspended in 1.5 mL fresh DMM containing no carbon source. The volume and OD of the cell suspension were measured and majority of the cell suspension was pipetted into a pre-weighed empty microcentrifuge tube and dried overnight in a 90°C heat block. The dry weight of the transferred cell suspension was calculated by subtracting the weight of the empty tube from dry weight of the cell suspension with the tube. Similarly, the dry weight of the medium was measured by evaporating fresh DMM containing no carbon source in pre-weighted microcentrifuge tubes. Finally, cell dry weight was calculated by dividing weight of dried cells (excluding weight of media

components) with volume and OD. Using this method, we determined the dry weight of *V. fischeri* cells is 0.451 ± 0.014 mg/OD/L.

The dry weight of protein and the amino acid compositions were measured following the protocol [138]. The dry weights of DNA and glycogen were measured using the kits from Sigma (Catalog Number DNAQF and MAK016). The dry weight of RNA was measured using the Qubit RNA HS Assay Kit. The DNA and RNA compositions were estimated using the sequence data. The dry weights of the lipids per gram of total cell dry weight were assumed to be the same as in the *E. coli i*JO1366 model and their compositions were estimated following the method used for in *i*JO1366. Dry weight of other biomass components were assumed to be the same as in the *i*JO1366 model. Sodium ions was added to the biomass due to a high intracellular sodium concentration in *Vibrio fischeri*.

AI.1.3 Growth-Phenotyping Experiments

Procedures for performing aerobic experiments were described previously [58]. For anaerobic experiments, inoculating fluid and sodium chloride were deoxygenated inside an anaerobic chamber for 48 hours prior to the experiments. Plates were deoxygenated in a sealed gas impermeable bag (Biolog) with oxygen absorber (Mitsubishi). The sealed plates were kept in the anaerobic chamber for 48 hours as well. The increase in turbidity was compared to a negative control with no added carbon source to determine whether cells grew. Additional anaerobic experiments with were performed in 17 by 100 mm test tubes in medium degassed by nitrogen to confirm growth of strains on a carbon source of interest, or detect strains with an intermediate increase in turbidity.

AI.1.4 Network Reconstruction and Modeling Simulation

Using the well curated *Escherichia coli i*JO1366 model as a template, we first constructed a draft model for *V. fischeri* ES114 by copying reactions from *i*JO1366 when we found homologous *V. fischeri* genes for *E. coli* genes responsible for these reactions. We defined the two genes homologous if they shared the same KEGG Orthology (KO) identifier, a unique number assigned to a group of homologous genes, and were the best reciprocal hits of each other. *E. coli* genes in gene-protein-reaction associations of the copied reactions were substituted with *V. fischeri* homologs. For any reaction associated with only a single gene, the reaction was not included in the draft model if a homologous *V. fischeri* gene was missing. In reactions with complex GPR associations with multiple subunits or isozymes, reactions were included if homologs for any isozyme or all subunits were present. In some cases where some subunits but not all were missing, reactions were included if the major catalytic subunit was present.

Flux balance analysis [116] that maximized the objective of biomass flux was used to predict the growth phenotypes of wild-type ES114 on a single carbon source minimal medium or gene essentiality in LBS medium. All calculations were performed using the CPLEX solver (IBM) in GAMS version 24.3.3. In addition to the exchange of the tested carbon source, exchange reactions with a non-zero lower limit were listed in Table AI.S1 to simulate the minimal medium or LBS medium compositions. A modified SMILEY algorithm [59] suggest gap-filled non-GPR reactions by adding a minimal set of *i*JO1366 reactions to the draft *V. fischeri* model that resolved the most model no-growth but experimental growth discrepancies. For gene essentiality predictions, the flux through the reaction(s) associated to the gene of interest was set to zero following GPR rules. A gene was defined as essential in LBS medium if the flux via the biomass reaction was zero.

We fit to our model the aerobic growth data of *V. fischeri* cell cultures in DMM containing different concentrations of glucose as a sole carbon source and estimated the growth-associated maintenance energy (GAM) to be 55.98 mmol ATP/gDW with the assumption that non-growth-associated maintenance energy (NGAM) was negligible. The GAM value is similar but slight higher than the value of 53.95 mmol ATP/gDW in *i*JO1366.

Potential infeasible energy generating cycles that involve energy metabolites or proton exchanges between periplasm and cytosol in the model were examined with a published protocol [139]. A dissipation reaction for an energy metabolite or proton was added to the metabolic network and flux through each reaction was maximized while carbon uptake was closed. If there existed the production of any of the energy metabolite, the reactions with non-zero fluxes were reviewed manually.

For calculating cell growth impacted by bioluminescence, we simulated a DMM medium environment containing *N*-acetyl-D-glucosamine as the sole carbon source. The lower limit of *N*-acetyl-D-glucosamine uptake rate was set at -2 mmol/gDW/h. Lower limits of medium components in DMM were all set at -20 mmol/gDW/h. The lower limit of oxygen uptake was chosen from three different levels (low -1 mmol/gDW/h, medium-5 mmol/gDW/h, high –10 mmol/gDW/h). Quantum yield was set at 10%, 20%, 50%, or 100% by changing the stoichiometric coefficients in the luciferase reaction (LUC). The stoichiometry of LUC also fixed the ratio between ATP consumed and photon produced at 7:1 [84]. FBA was used to maximize the growth rate at varying luciferase flux, quantum yield, and oxygen uptake values. Finally we normalize this number with the maximum growth rate at zero bioluminescence flux.

AI.2 Figures and Tables

Table AI.S1 Medium Recipe for *In Silico* Simulation

Reactions for simulating *V. fischeri* DMM		**Additional exchange reactions for simulating LBS medium**			
Reaction Abbreviation	Reaction Name	Reaction Abbreviation	Reaction Name	Reaction Abbreviation	Reaction Name
EX_K_E	K^+ exchange	EX_ADN_E	Adenosine exchange	EX_INS_E	Inosine exchange
EX_CA2_E	Calcium exchange	EX_ALA-L_E	L-Alanine exchange	EX_LEU-L_E	L-Leucine exchange
EX_CBL1_E	Cob(I)alamin exchange	EX_AMP_E	AMP exchange	EX_LIPOATE_E	Lipoate exchange
EX_CL_E	Chloride exchange	EX_ARG-L_E	L-Arginine exchange	EX_LYS-L_E	L-Lysine exchange
EX_CO2_E	CO_2 exchange	EX_ASN-L_E	L-Asparagine exchange	EX_MET-L_E	L-Methionine exchange
EX_COBALT2_E	Co^{2+} exchange	EX_ASO3_E	Arsenite exchange	EX_NAC_E	Nicotinate exchange
EX_CU2_E	Cu^{2+} exchange	EX_ASP-L_E	L-Aspartate exchange	EX_PHE-L_E	L-Phenylalanine exchange
EX_FE2_E	Fe^{2+} exchange	EX_BTN_E	Biotin exchange	EX_PHEME_E	Protoheme exchange
EX_FE3_E	Fe^{3+} exchange	EX_CD2_E	Cadmium exchange	EX_PNTO-R_E	(R)-Pantothenate exchange
EX_H_E	H^+ exchange	EX_CMP_E	CMP exchange	EX_PRO-L_E	L-Proline exchange
EX_H2O_E	H_2O exchange	EX_CYS-L_E	L-Cysteine exchange	EX_PYDX_E	Pyridoxal exchange
EX_MG2_E	Mg exchange	EX_DAD-2_E	Deoxyadenosine exchange	EX_RIBFLV_E	Riboflavin exchange
EX_MN2_E	Mn^{2+} exchange	EX_DCYT_E	Deoxycytidine exchange	EX_SER-L_E	L-Serine exchange
EX_MOBD_E	Molybdate exchange	EX_FOL_E	Folate exchange	EX_SKM_E	Shikimate exchange
EX_NA1_E	Sodium exchange	EX_GLN-L_E	L-Glutamine exchange	EX_THM_E	Thiamin exchange
EX_NH4_E	Ammonia exchange	EX_GLU-L_E	L-Glutamate exchange	EX_THR-L_E	L-Threonine exchange
EX_NI2_E	Ni^{2+} exchange	EX_GLY_E	Glycine exchange	EX_THYMD_E	Thymidine exchange
EX_NO3_E	Nitrate exchange	EX_GMP_E	GMP exchange	EX_TRP-L_E	L-Tryptophan exchange
EX_PI_E	Phosphate exchange	EX_GSN_E	Guanosine exchange	EX_TYR-L_E	L-Tyrosine exchange
EX_SEL_E	Selenate exchange	EX_H2S_E	Hydrogen sulfide exchange	EX_UMP_E	UMP exchange
EX_SLNT_E	Selenite exchange	EX_HG2_E	Hg^{2+} exchange	EX_URA_E	Uracil exchange
EX_SO4_E	Sulfate exchange	EX_HIS-L_E	L-Histidine exchange	EX_URI_E	Uridine exchange

Reactions for simulating *V. fischeri* DMM		**Additional exchange reactions for simulating LBS medium**			
Reaction Abbreviation	Reaction Name	Reaction Abbreviation	Reaction Name	Reaction Abbreviation	Reaction Name
EX_TUNGS_E	tungstate exchange	EX_HXAN_E	Hypoxanthine exchange	EX_VAL-L_E	L-Valine exchange
EX_ZN2_E	Zinc exchange	EX_ILE-L_E	L-Isoleucine exchange		
EX_O2_E	O_2 exchange (for aerobic conditions only)				

Vibrio fischeri Transcriptomic Experiments and Analysis Supplementary Materials

AII.1 Methods

AII.1.1 Collection of Bacterial RNA (1st Dataset)

Bacterial RNA was obtained from V. fischeri cells either grown in SWT to an OD600 of ~0.5 (‘cultured’), or cultured in SWT to an OD600 of ~0.5, then incubated in artificial seawater for 18 h (‘planktonic’), as described above (Fig. 1A). In both cases, the bacteria were collected by centrifugation at 12 000 rpm for 2 min (Model 5254R, Eppendorf) immediately prior to RNA extraction and purification using a ZR RNA MicroPrep Kit (Zymo Research). To obtain squid-associated bacteria, freshly hatched juvenile squid from the University of Wisconsin-Madison aquarium facility were collected into V. fischeri-free IO. Animals were subsequently exposed for 3 h to approximately 5000 CFU of V. fischeri per mL, and then transferred to fresh *V. fischeri*-free IO. Animals were maintained on a 12:12 hour day: night cycle in batches of approximately 30 to 35 animals to optimize uniformity of colonization timing between individual animals. After ~36 h, colonization of the animals was verified by determining their production of bioluminescence using a luminometer (Turner TD 20/20). At approximately 48 h post-colonization, equivalent to the end of night time, animals were collected in batches of 100 into 80 mL of filter-sterilized IO, and exposed to light for 15 min to induce the normal dawn-cued expulsion of their bacterial population. Squid were subsequently removed, and the expelled bacteria were collected by centrifugation at 12 000 rpm for 10 min at room temperature in an SS-34 rotor (Thermo Fisher Scientific). The supernatant was discarded, and total bacterial RNA was immediately extracted and purified using a ZR RNA MicroPrep Kit. In all cases, the time from initial light exposure to RNA extraction was<30 min.

AII.1.2 RNA-Seq library preparation and sequencing (1st Dataset)

Three replicate RNA samples for each condition (squid associated, planktonic, or cultured) were processed for RNA-Seq transcriptomic analysis. Starting with total RNA, ribosomal RNA was removed using the Ribo-Zero Gold Epidemiology Kit (Epicentre), which removed prokaryotic (*V. fischeri*) rRNA as well as any contaminating eukaryotic *(E. scolopes*) rRNA. The resulting ribo-depleted RNA was then used for library preparation using the TruSeq RNA sample preparation kit (Illumina). Libraries were sequenced using Illumina HiSeq 2500 with high output, V4 chemistry, and 100-bp single-end reads. Fastq files were de-multiplexed, yielding on average 18M reads per sample (range: 5 to 26M reads per sample). Per-sample raw fastq files and processed CDS and rRNA count tables have been submitted to NCBI Gene Expression Omnibus (GEO) with accession number GSE80607.

AII.1.3 Sequence Read Processing and Mapping (1st Dataset)

Single-end fastq files were trimmed with Trimmomatic v.0.33. Reads from each sample were mapped to the *V. fischeri* ES114 genome (NC_006840.2) using bwa v.0.6.2 (Burrows–Wheeler Aligner), applying commands index, aln, and samse. The resulting SAM files were processed with SAMtools v.0.1.19 to generate BAM files. The numbers of reads mapping to protein-coding (CDS) or rRNA genes were calculated using the htseq-count command of HTSeq v.0.6.1p1. Ribodepleted samples for the primary three-treatment comparison yielded an average of 10.0M CDS reads mapped per sample; ribo-depleted samples for the low-biomass comparison yielded an average of 5.3M CDS reads mapped per sample; in addition, ~34 000 rRNA reads were mapped per sample across both sets of ribo-depleted samples.

AII.1.4 Differential Expression Analysis (1st Dataset)

Detection of genes differentially expressed (i.e. relative transcript abundance) between conditions was performed using the R packages DESeq2 and NOISeq from the Bioconductor program. Low counts were removed using the NOISeq command filtered. Data. After loading data into a data frame, DESeq2 was run. The three conditions (squid-associated, planktonic and cultured) were contrasted pairwise for all genes, and results exported as Benjamini–Hochberg adjusted p-values and log2 FC. Differentially expressed genes were identified using tiered cutoffs of these values, with the most stringent cutoff being an adjusted p-value<0.001 and abs(log2 FC)>3.0 (three replicates per condition). All code used in data analysis and manuscript preparation is available at https://github.com/cuttlefishh/papers/tree/master/ vibrio-fischeri-transcriptomics.

AII.1.5 Flux-Coupling Analysis of Symbiont Sugar Uptake (1st Dataset)

We analyzed the transcriptomic data using a constraint-based genome-scale metabolic model developed for *V. fischeri* ES114. Flux-coupling analysis was performed on the model to identify which reactions (and their associated genes) are fully, partially, or directionally coupled to nutrient uptake [13]. These types of reaction–nutrient coupling indicate that if the coupled reaction is active (i.e. carries a non-zero flux) then the nutrient must be consumed. As a result, the genes associated with these coupled reactions might serve as biomarkers for nutrient uptake. Flux-coupling analysis was performed assuming all sugars, amino acids, nucleotides and inorganic nutrients were available (by making the lower limits for all the exchange reactions in the model negative). The genes associated with reactions coupled to only one nutrient were compared with the list of differentially regulated genes to allow us to determine which nutrients could be used by

V. fischeri.

AII.1.6 Bacterial Strains and Growth Conditions (2nd Dataset)

Wild-type *V. fischeri* ES114 and GlcNAc EIIBC transporter mutant Δ*nagE* were grown overnight at 28°C with shaking at 225 rpm in 12 by 75 mm culture tubes containing 3ml Luria-Bertani salt (LBS) medium [137]. The overnight cultures were subcultured into either SWTO [140] or SWTO with 20 mM GlcNAc with an initial OD of 0.05. Every condition was run in triplicates. The cultures were harvested at 0.5 for RNA extraction and subsequent RNA extraction.

AII.1.7 RNA Isolation, Sequencing, and Mapping (2nd Dataset)

All procedures were performed in an RNAase free environment treated with RNase AWAY Surface Decontaminant (Thermo Scientific).Total RNA from harvested cells was extracted at room temperature from the aqueous phase using the RNeasy Mini Kit (Qiagen), RNase-Free DNase (Qiagen), RNase-free water (Ambion), following manufacturer's protocols. The quantity and quality of total RNA was assessed using a NanoDrop spectrophotometer. Samples with a 260/280 ratio greater than 1.9 were selected for sequencing.

After sequencing, raw reads were processed to exclude low quality reads and the quality of reads were confirmed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). High-quality reads were mapped to the *V. fischeri* ES114 genome with the software package Bowtie (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml).

AII.1.8 Gene Set Enrichment Analysis (2nd Dataset)

We followed a protocol of gene set enrichment analysis (GSEA) [91] to interpreting gene expression data. Gene sets were downloaded from BioCyc *V. fischeri* ES114 database version 22.0 [61]. We used 264 out of all 374 gene sets that have a size greater

than two. Gene sets were ranked according to log2 fold changes (log2 FCs) calculated from EdgeR [141]. GSEA was performed on these 264 gene sets using the Python package, GSEAPY (https://github.com/BioNinja/GSEApy).

AII.1.9 Modeling Acetate and Ammonium Secretion (2nd Dataset)

The genome-scale metabolic model of *V. fischeri* ES114, *i*VF846, and flux balance analysis [116] were used to calculate acetate and ammonium secretion rates at the maximum growth rate under a specified media condition (Appendix II, Table AII.S1). Sugar concentrations were varied to achieve a broad range of maximum growth rates. The anaerobic or medium level oxygen uptake conditions were simulated by setting the lower limit of the oxygen exchange flux to 0 and -5 mmol/gDW/h respectively. In the case where alternate optimal solutions exist at a fixed sugar and oxygen uptake levels, the minimum and maximum values of the alternate optimal solutions were calculated using flux variability analysis.

AII.2 Figures and Tables

Table AII.S1 Medium Recipe for Acetate and Ammonium Secretion Simulation

Reaction Abbreviation	Reaction Name	Lower Limit (mmol/gDW/h)	Upper Limit (mmol/gDW/h)
EX_K_E	K^+ exchange	-20	1000
EX_CA2_E	Calcium exchange	-20	1000
EX_CBL1_E	Cob(I)alamin exchange	-20	1000
EX_CL_E	Chloride exchange	-20	1000
EX_CO2_E	CO_2 exchange	-20	1000
EX_COBALT2_E	Co^{2+} exchange	-20	1000
EX_CU2_E	Cu^{2+} exchange	-20	1000
EX_FE2_E	Fe^{2+} exchange	-20	1000
EX_FE3_E	Fe^{3+} exchange	-20	1000
EX_H_E	H^+ exchange	-20	1000
EX_H2O_E	H2O exchange	-20	1000
EX_MG2_E	Mg exchange	-20	1000
EX_MN2_E	Mn^{2+} exchange	-20	1000
EX_MOBD_E	Molybdate exchange	-20	1000
EX_NA1_E	Sodium exchange	-20	1000
EX_NH4_E	Ammonia exchange	-20	1000
EX_NI2_E	Ni^{2+} exchange	-20	1000
EX_NO3_E	Nitrate exchange	-20	1000
EX_PI_E	Phosphate exchange	-20	1000
EX_SEL_E	Selenate exchange	-20	1000
EX_SLNT_E	Selenite exchange	-20	1000
EX_SO4_E	Sulfate exchange	-20	1000
EX_TUNGS_E	tungstate exchange	-20	1000
EX_ZN2_E	Zinc exchange	-20	1000
EX_O2_E	O_2 exchange	0, -5, or -10	1000
EX_GLC_E	D-Glucose	varying	1000
EX_ACGAM_E	N-Acetyl-D-glucosamine exchange	varying	1000
EX_ACGAM2_E	Chitobiose exchange	varying	1000
EX_GAM_E	D-Glucosamine exchange	varying	1000
EX_GLYC_E	Glycerol exchange	varying	1000

MEGS Supplementary Materials

AIII.1 Methods

AIII.1.1 *In Silico* Modeling

A newly constructed genome-scale metabolic model of *V. fischeri* strain ES114, *i*VF846, was used in this work. Reactions and metabolites from an *Escherichia coli* model, iJO1366, [115] were transferred into the draft model of iVF846 when orthologs to *E. coli* metabolic genes were found in *V. fischeri*. By our definition, the orthologous genes shared the KEGG ortholog identifier and were best reciprocal hits in the KEGG Sequence Similarity Database (SSDB). The model contains 846 genes, 907 metabolites, and 1509 reactions. Flux-balance analysis (FBA) [116] was used to calculate the growth rate of *V. fischeri* by maximizing flux through a defined biomass objective function. In FBA, the minimal or rich medium was simulated by giving negative values to the lower limits for the exchange fluxes of the medium components in the model. A carbon source was predicted to be a sole carbon source if the FBA-predicted growth rate was greater than zero in minimal medium supplemented with this carbon source. To simulate a knockout strain, the fluxes of metabolic reactions associated with the gene were fixed at zero, unless isozymes were present. A gene was predicted as essential if the FBA-predicted growth rate of the strain containing a single gene deletion was zero. A modified version of the mixed-integer linear programming algorithm, SMILEY [59], was used to predict missing metabolic genes (and their associated reactions) when the model incorrectly predicted wild-type or mutant *V. fischeri* strains cannot grow in a particular medium. The algorithm was modified from its original implementation by minimizing the total number of metabolic

genes (instead of reactions) that need to be added from iJO1366 [115] (instead of KEGG) to the *V. fischeri* model to enable growth and reconcile false predictions.

AIII.1.2 Strain Construction

Wild-type *E. coli* BW25113, and wild-type *V. fischeri* ES114 were used in this work. *E. coli* knockout strains (derived from BW25113) containing kanamycin (Kan) resistance genes were obtained from the Keio collection (Open Biosystems) [45,46]. The temperature-sensitive plasmid pCP20 was used to remove the *kan* gene from the mutants as described previously [142]. The resulting kanamycin sensitive *E. coli* knockout strains were used as recipient strains for the *V. fischeri* genomic library. Knockout mutants of *V. fischeri* ES114 were constructed using conjugation and homologous recombination as described previously [72,143–145]. To construct ΔVF_0892, 10 mM pantothenate and 10 mM β-alanine were supplemented in the LBS growth medium.

AIII.1.3 Plasmid Construction for Single-Gene Complementation

For *E. coli* complementation experiments, a single *V. fischeri* gene was cloned into the multiple cloning site of pZE21MCS (EXPRESSYS) using Gibson cloning. The construct was transformed into the corresponding *E. coli* knockout strains and colonies were selected on LB agar containing 50 µg kanamycin per mL. For *V. fischeri* complementation experiments, a single *V. fischeri* gene was cloned into the multiple cloning site of pVSV105 [146] using Gibson cloning. These plasmids containing a *V. fischeri* gene were introduced into *V. fischeri* ES114 knockout strains by conjugation.

AIII.1.4 Growth Conditions and Complementation Experiments

Unless otherwise noted, *E. coli* strains were grown at 37°C and *V. fischeri* strains were grown at 28°C, both with shaking at 225 rpm. *E. coli* strains were grown in Luria-Bertani (LB) or a morpholinepropanesulfonic acid (MOPS)-buffered minimal medium [147]. Because *E. coli* grows poorly on glutamine as a sole carbon source, vitamin

supplements (0.05 mM thiamine, 0.05 mM niacinamide, and 20 nM biotin) were added to the minimal medium to shorten the experiments in which the glutamine transporter is complemented. *V. fischeri* strains were grown in Luria-Bertani salt (LBS) [113] or *V. fischeri* defined minimal medium (DMM). Overnight cultures of ΔVF_0892, were grown in LBS supplemented with 10 mM pantothenate and 10 mM β-alanine. When appropriate, 50 µg kanamycin or 5 µg chloramphenicol per mL was added to the media. For complementation experiments, an overnight LB (or LBS) culture of each strain was subcultured by ~1:100 dilution into fresh minimal medium with a starting optical density of 0.02 at 600 nm (OD_{600}). The OD_{600} of the culture in a 96-well plate was measured by an Infinite M200 plate reader (Tecan) every 15 minutes with 3 mm orbital shaking. In complementation experiments with a VF_0892 plasmid in Δ*panD* or ΔVF_0892, an overnight LB (or LBS) culture of each strain was washed twice in minimal medium and subcultured by 1:100 dilution into fresh minimal medium. Once the subculture grew to about mid-exponential phase, it was washed twice and subcultured again into fresh minimal medium for the growth curve measurements. All experiments testing a VF_A0062 plasmid were performed at 28°C, since pZEVFA0062 did not complement Δ*mtlD* well when grown at 37°C.

AIII.1.5 Growth-Phenotyping Experiments

To test sole carbon sources of *V. fischeri*, *V. fischeri* strain ES114 was grown in inoculating fluid supplemented with 255 mM sodium chloride, on PM1 or PM2A plates containing single carbon sources (Biolog) according to the manufacturer's protocol. The plates were incubated at 28°C in an OmniLog incubator-reader (Biolog) and turbidity was measured every 15 minutes for 48 hours. The increase in turbidity was compared to a negative control with no added carbon source to determine whether cells grew.

Additional experiments were performed in 17 by 100 mm test tubes to confirm growth of strains on a carbon source of interest, or detect strains with an intermediate increase in turbidity.

AIII.1.6 Genomic Library Construction

The genomic DNA of *V. fischeri* ES114 strain was extracted from LBS culture using the DNeasy Blood & Tissue Kit (Qiagen). The extracted DNA was then fragmented at 10% amplitude for 5 seconds using the Sonic Dismembrator Model 500 (Fisher Scientific). DNA fragments between ~2 and ~5 kbp were size-selected from a 1% agarose gel in Tris-acetate-EDTA buffer and purified with the Zymoclean Gel DNA Recovery Kit (Zymo). The DNA fragments were ligated into the HinCII site of the multiple cloning site of the pZE21MCS1 vector and transformed into 50 µL of *E. coli* MegaX suspension (Invitrogen) following the protocol of Forsberg *et al.* [112]. The average insert size of the library was ~2.3 kbp and the titer of the library was 22 million colony forming units (CFUs). Transformed cells were transferred to 10 mL of LB containing 50 µg of kanamycin per mL and grown overnight. The overnight culture was used to extract the library of plasmids using the QIAprep Spin Miniprep Kit (Qiagen).

AIII.1.7 Gene Selection from a *V. fischeri* Genomic Library

The *V. fischeri* genomic library was transformed into competent cells of an *E. coli* recipient strain and recovered in 1 mL SOC for an hour. The recovered cells were then pelleted at 6000 rpm for 3 min, and transferred to 50 mL of selective medium with 50 µg kanamycin per mL to grow at 37°C. After the optical density at 600 nm (OD_{600}) reached 1, the cells were subcultured into 50 mL of fresh selective medium. After the OD_{600} of the cells reached 1 again, they were plated on LB plus kanamycin plates. Single colonies were picked to confirm growth in selective medium and the first and last

700 bp of the plasmid inserts were subsequently sequenced to identify the *V. fischeri* gene(s) included in the plasmid.

AIII.1.8 Expression and Purification of Decarboxylases

E. coli panD (b0131), *V. fischeri panP* (VF_0892), and *V. fischeri gadA* (VF_1064) were amplified from the genomic DNA using Phusion High-Fidelity DNA Polymerase (NEB). The PCR fragments containing VF_0892 were digested with NcoI and XhoI and cloned into pET28 (Novagen). The resulting plasmid pETVF0892 was transformed into *E. coli* X90(DE3) competent cells [148]. The PCR fragments containing b0131 or VF_1064 were inserted into the pET28 vector with an N-terminal His-tag sequence followed by a TEV site. The resulting plasmids pETb0131 and pETVF1064 were transformed into *E. coli* BL21 (DE3) competent cells. For expression, an inoculum was started in LB medium plus 50 µg kanamycin per mL from an overnight culture. The cells were grown at 28°C until reaching an OD_{600} of 0.6. Then cells were induced by 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) overnight at 18°C. The cell lysate was extracted from the collected cells by BugBuster Master Mix (EMD Millipore) following the manufacturer's protocol. The lysate was incubated with HisPur Ni-NTA Resin (Thermo Scientific) for 1 hour at 4°C and passed through a Pierce Centrifuge Column (Thermo Scientific). The column was washed in 50 mM NaH_2PO_4, 300 mM NaCl, and 20mM imidazole (pH 8), and the VF_0892 protein was eluted in 50 mM NaH_2PO_4, 300 mM NaCl, and 250 mM imidazole (pH 8). The eluted protein was dialyzed against 100 mM Tris-HCl (pH 7.5), and concentrated with an Amicon Ultra Centrifugal Filter Unit (EMD Millipore). The final products were analyzed with SDS-PAGE gels and their concentrations were determined with a Bicinchoninic Acid Kit (Sigma). The eluted protein was dialyzed in 20 mM Tris-HCl to reduce the amount of solutes that went through the GC-MS column. *E.*

coli X90(DE3) containing the empty pET28b(+) vector was subjected to the same expression and purification procedures, and the product was used as a control in detecting *V. fischeri* PanP activity using GC-MS.

AIII.1.9 Detection of β-alanine from *In Vitro* Enzyme Assays

The *V. fischeri* PanP activity was detected by measuring the formation of β-alanine in five different conditions, each with a 200 µL reaction volume. All reaction mixtures contained 20 mM Tris-HCl, 5 mM $MgSO_4$, and 750 µM pyridoxal 5'-phosphate (pH 7.5). In addition to the buffer components, the first condition also contained 36 µg of purified *V. fischeri* PanP and 50 nmols of aspartate as the substrate. The second condition contained 36 µg purified *V. fischeri* PanP, but no substrate. The third condition contained 50 nmols of aspartate, but no protein. The fourth condition contained 36 µg purified *V. fischeri* PanP that had been heat-inactivated at 90°C for 30 min, and 50 nmols of aspartate. The last condition contained 20 µg protein obtained from the empty vector lysate and 50 nmols of aspartate. The reaction mixtures were incubated at 37°C for 15 hours, and then stopped by heat inactivation at 90°C for 30 min. The second and the fourth conditions were tested in biological duplicates and the other conditions were tested in biological triplicates. The heat-inactivated reaction mixtures were spun down at 15,000 rpm for 3 minutes, and their supernatants were taken for quantitation. A uniformly labeled, [U-13C] β-alanine internal standard (Cambridge Isotope Laboratories) was used for quantification via an isotope-ratio method using gas chromatography-mass spectrometry (GC-MS) [138]. Prior to sample quantification, a suitable β-alanine fragment formula was identified. This process includes (i) analyzing the mass spectrum of an unlabeled β-alanine sample to propose a feasible structure for the molecular ion of interest, (ii) comparing the theoretical and measured isotopic distributions of said structure, and (iii) verifying that the number of

carbon backbone atoms present on the derivatized structure matches the base peak mass shift predicted to be seen in the [U-13C] β-alanine mass spectrum relative to the unlabeled spectrum [149]. Once a fragment for quantification was identified, the purity (i.e., extent of labeling) of labeled β-alanine was measured and subsequently used to correct for the unlabeled portion of the labeled standard in downstream calculations. Following this step, aliquots of labeled internal standard were quantified using a known amount of unlabeled standard and calculating the ratio of 12C/13C after correcting both for natural abundances of other isotopes present in the derivatized fragment, and for the unlabeled portion of the labeled standard, using the freely available software, IsoCor [150]. Once characterized, known amounts of labeled standard were mixed with the supernatant of the enzymatic assay and dried at 90°C. This step was followed by derivatization in a volumetric 1:1 ratio of pyridine with with *N*-tert-butyldimethylsilyl-*N*-methyltrifluoroacetamide plus 1% tert-butyldimethylchlorosilane (MTBSTFA with 1% t-BDMCS) at 90°C for 30 minutes to confer thermal stability and increased volatility amenable for analysis on the GC-MS instrument. Derivatized samples were centrifuged at 15,000 rpm for 3 minutes to sediment insoluble material, producing a cleaner supernatant for injection onto the GC-MS. Samples were run on a single quadrupole GCMS -QP2010S (Shimadzu) in electron ionization mode with a Rtx 5ms (Restek) low-bleed, fused-silica column for separation with helium as a carrier gas operating under a linear velocity control mode with a split ratio of 0.50, and a column flow of 1.50 mL/minute. The temperature program for separation of β-alanine began with holding the column oven temperature at 35°C for 10 minutes, ramping up at 25°C/minute to 300°C, and holding for 19.4 minutes. Operational parameters included an injection temperature of 240°C, ion

source temperature of 260°C, interface temperature of 240°C, and a mass scan range of 100-450 m/z. To test the detection limit of our β-alanine quantification method, the method was used to measure a known amount of unlabeled β-alanine. Each sample (n=3) contained 0.125, 0.25, 0.5, 1, or 2 nmols of unlabeled β-alanine. The method was able to detect the presence of unlabeled β-alanine in all these samples. Therefore, the detection limit of the method is below 0.125 nmol.

AIII.1.10 DC-PEPC-MDH-Linked Assays

Decarboxylase (DC), phosphoenolpyruvate carboxylase (PEPC), and malate dehydrogenase (MDH) -linked assays were performed by mixing 80 µL freshly prepared mixture A and 120 µL mixture B. Mixture A contained 100 mM Tris-HCl, 10 mM $MgSO_4$, 1 mM PLP, and various concentrations of aspartate and glutamate (pH 8.05). Mixture B contained purified decarboxylase and Infinity Carbon Dioxide Liquid Stable Reagent (Infinity, Thermo Fisher Scientific). The final decarboxylase concentrations used were 54 µM PanD, 35 µM VF_1064 enzyme, 28 µM *V. fischeri* PanP for assays at 37°C, and 14 µM *V. fischeri* PanP for assays at 28°C. Like similar DC-PEPC-MDH-linked assays [151–155], our assays were carried out in air. The interference of exogenous CO_2 from air and buffer solution was accounted for using a negative control, in which no substrate was added. The signal produced by the negative control was linear during the measurement period, and was subtracted from all signals produced by other samples containing substrates. Values of K_m and k_{cat} were determined from nonlinear fitting into the Michaelis-Menten equation using KaleidaGraph (Synergy Software). Averages across the 3 technical replicates were used as data points while standard deviations for each data point were used as weights during the curve fitting. The K_m and k_{cat} values are reported with standard errors.

AIII.1.11 qPCR Analysis

Wild-type *V. fischeri* cells were harvested at an OD_{600} of ~0.6 in DMM supplemented with glucose or mannitol as sole carbon source. Total RNA from the cell pellets was harvested using the Quick-RNA MicroPrep Kit (Zymo Research) with a 15-min on column DNase treatment. RNA concentration and quality were assessed on a NanoDrop (Thermo Scientific) spectrophotometer. RT for synthesizing cDNA and real-time qPCR were performed using the GoTaq 2-Step RT-qPCR System (Promega) and an AriaMx Real-time PCR machine (Agilent). The differential expression of VF_A0062 (forward primer, TGGATATTCCGGGTGGTAAA, reverse primer, ACGGGTCTTGTTCTGCAAGT) in the mannitol minimal medium and the glucose minimal medium was normalized to the control gene (*V. fischeri* polA, VF_0074, forward primer, CGACAGCAGCAGAAGTGAAG, reverse primer, AGCAAGACCAAACGCACTC) using the GED formula [156].

AIII.1.12 Squid Colonization Competitions

Freshly hatched juvenile squid were collected from the rearing facility at the University of Hawaii and placed in filter-sterilized seawater. Squid were exposed to an approximately 1:1 mixed population of two strains consisting of *V. fischeri* ES114 carrying a chromosomally inserted erythromycin-GFP marker [55], and the indicated mutant strain, at a total of 3000-5000 CFU/mL. Squid were incubated with bacteria for 3 hours, then transferred to individual vials of *V. fischeri*-free seawater for an additional 18-21 hours. Colonized squid were subsequently anesthetized on ice and placed at -80°C for surface sterilization. Individual squid were then homogenized and plated on LBS and LBS + erythromycin agar plates as described previously [67], and the ratio of strains present in the light organ was determined by counting the unmarked and erythromycin marked

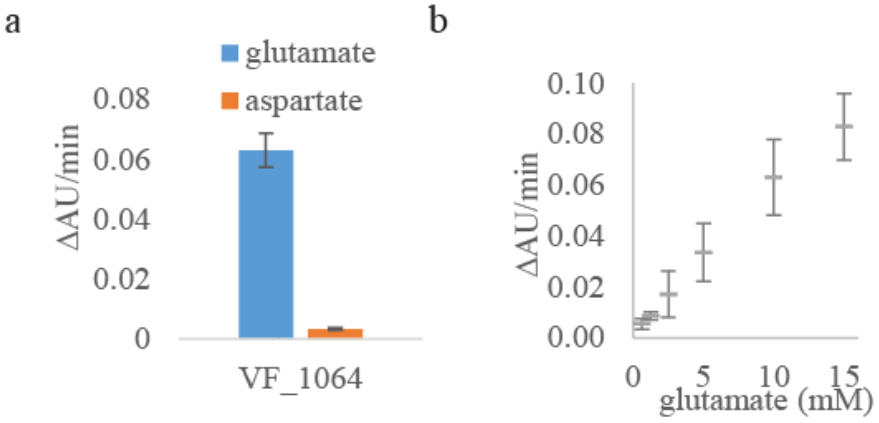
colonies. The relative competitive index (RCI) of the co-colonizing strains was calculated as follows: RCI = log(CFU mutant / CFU wild type)/(inoculum CFU mutant / inoculum CFU wild type).

AIII.2 Figures and Tables

Table AIII.S2 Recipient Strain and Selective Medium Design.

Draft model and experiment discrepancies			*E. coli* genes (bnums) missing *V. fischeri* ortholog	Recipient strain	Selective medium
Category	Model prediction	Experimental data			
Growth phenotype	No growth in minimal medium	Growth in minimal medium	*panD* (b0131)	Δ*panD*	MOPS minimal medium + 20 mM glucose
	Mannitol is not a sole carbon source	Mannitol is a sole carbon source	*mtlD* (b3600)	Δ*mtlD*	MOPS minimal medium + 20 mM mannitol
	N-acetyl-neuraminate is not a sole carbon source	N-acetyl-neuraminate is a sole carbon source	*nanT* (b3224)	Δ*nanT*	MOPS minimal medium + 20 mM N-acetylneuraminate
Gene essentiality	Glutamine synthase (VF_0098) is essential in LBS medium	Glutamine synthase (VF_0098) is non-essential in LBS medium	*glnH* (b0811) *glnP* (b0810) *glnQ* (b0809)	Δ*glnP*	MOPS minimal medium + 20 mM glutamine + vitamin supplement (0.05 mM thiamine, 0.05 mM niacinamide, and 20 nM biotin)

Figure AIII.S2 Substrate Specificity of the VF_1064 Enzyme. Average reaction rates (ΔAU/min) are shown for three technical replicates of the DC-PEPC-MDH-linked assays performed at the specified aspartate or glutamate concentration with standard deviation shown as the error bars. (a) Reaction rates of the VF_1064 enzyme with a concentration of 10 mM of either glutamate or aspartate at 37°C. (b) Reactions rates of the VF_1064 enzyme with various concentrations of glutamate at 37°C.



Zymomonas mobilis Metabolic Discoveries Supplementary Materials

AIV.1 Methods

AIV.1.1 Bacterial Strains and Growth Conditions

E. coli BW25113 was obtained from *E. coli* genetic stock center. The *E. coli* knockout strains*ubiC::kan*, *pdxB::kan*, *bioH::kan*, *frdA::kan*, *frdB::kan*, *frdC::kan*, *frdD::kan*, *fumA::kan, fumC::kan*, *fumB::kan*, *pabA::kan, pabB::kan*, and *serC::kan* were obtained from the Keio collection (Open Biosystems) [157]. To generate mutants with multiple gene deletions, the kanamycin resistance gene (*kan*) was first removed using the temperature sensitive pCP20 plasmid as described previously [142]. Additional gene deletions were incorporated using P1 transduction [158] with the appropriate kanamycin resistant mutants as donor strains followed by selection on Luria-Bertani (LB) agar plates with kanamycin (50 ug/mL). The *kan* removal and P1 transduction steps were used to create *E. coli ΔfumAC*, and *ΔfumACB* knockout mutants. *Z. mobilis* ZM4 was obtained from ATCC (ATCC 31821). The *Z. mobilis* Tn5 mutants were obtained from the *Z. mobilis* ZM4 Tn5 mutant collection generated by Skerker et al. [49]. *E. coli* strains were grown at 37°C with shaking and *Z. mobilis* strains were grown at 30°C without shaking. *E. coli* strains were grown aerobically in LB broth or selective medium. *Z. mobilis* strains were grown in *Zymomonas* Rich Medium Glucose (ZRMG) and *Zymomonas* Minimal Medium Glucose (ZMMG).

Bibliography

1. Edwards JS, Palsson BO: **Systems properties of the *Haemophilus influenzae* Rd metabolic genotype**. *J. Biol. Chem.* 1999, **274**:17410–17416.

2. Zhang C, Hua Q: **Applications of genome-scale metabolic models in biotechnology and systems medicine**. *Front. Physiol.* 2016, **6**.

3. King ZA, Lloyd CJ, Feist AM, Palsson BO: **Next-generation genome-scale models for metabolic engineering**. *Curr. Opin. Biotechnol.* 2015, **35**:23–29.

4. O'Brien EJ, Monk JM, Palsson BO: **Using genome-scale models to predict biological capabilities**. *Cell* 2015, **161**:971–987.

5. McCloskey D, Palsson BO, Feist AM: **Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli***. *Mol. Syst. Biol.* 2014, **9**:661–661.

6. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation.** *Nat. Rev. Genet.* 2006, **7**:130–41.

7. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR).** *Genome Biol.* 2003, **4**:R54.

8. Kim J, Reed JL: **OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.** *BMC Syst. Biol.* 2010, **4**:53.

9. Patil KR, Rocha I, Förster J, Nielsen J: **Evolutionary programming as a platform for *in silico* metabolic engineering.** *BMC Bioinformatics* 2005, **6**:308.

10. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.** *Biotechnol. Bioeng.* 2003, **84**:647–57.

11. Pharkya P, Burgard AP, Maranas CD: **OptStrain : A computational framework for redesign of microbial production systems**. 2004, doi:10.1101/gr.2872004.14.

12. Ranganathan S, Suthers PF, Maranas CD: **OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions.** *PLoS Comput. Biol.* 2010, **6**:e1000744.

13. Burgard AP, Nikolaev E V, Schilling CH, Maranas CD: **Flux coupling analysis of genome-scale metabolic network reconstructions.** *Genome Res.* 2004, **14**:301–12.

14. Tervo CJ, Reed JL: **FOCAL: an experimental design tool for systematizing metabolic discoveries and model development.** *Genome Biol.* 2012, **13**:R116.

15. Tervo CJ, Reed JL: **Expanding metabolic engineering algorithms using feasible space and shadow price constraint modules.** *Metab. Eng. Commun.* 2014, **1**:1–11.

16. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat. Protoc.* 2010, **5**:93–121.

17. Hamilton JJ, Reed JL: **Software platforms to facilitate reconstructing genome-scale metabolic networks**. *Environ. Microbiol.* 2014, **16**:49–59.

18. Orth JD, Palsson B: **Systematizing the generation of missing metabolic**

knowledge. *Biotechnol. Bioeng.* 2010, **107**:403–412.

19. Thiele I, Vlassis N, Fleming RMT: **FASTGAPFILL: Efficient gap filling in metabolic networks**. *Bioinformatics* 2014, **30**:2529–2531.

20. Hartleb D, Jarre F, Lercher MJ: **Improved metabolic models for *E. coli* and *Mycoplasma genitalium* from GlobalFit, an algorithm that simultaneously matches growth and non-growth data sets**. *PLOS Comput. Biol.* 2016, **12**:e1005036.

21. Prigent S, Frioux C, Dittami SM, Thiele S, Larhlimi A, Collet G, Gutknecht F, Got J, Eveillard D, Bourdon J, et al.: **Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks.** *PLoS Comput. Biol.* 2017, **13**:e1005276.

22. Frioux C, Schaub T, Schellhorn S, Siegel A, Wanko P: **Hybrid metabolic network completion**. In *LPNMR 2017: Logic Programming and Nonmonotonic Reasoning*. 2017:308–321.

23. Hosseini Z, Marashi S-A: **Discovering missing reactions of metabolic networks by using gene co-expression data**. *Sci. Rep.* 2017, **7**:41774.

24. Liu L, Zhang Z, Sheng T, Chen M: **DEF: an automated dead-end filling approach based on quasi-endosymbiosis**. *Bioinformatics* 2016, **33**:btw604.

25. Ganter M, Kaltenbach H-M, Stelling J: **Predicting network functions with nested patterns**. *Nat. Commun.* 2014, **5**:3006.

26. Zhang M, Cui Z, Oyetunde T, Tang Y, Chen Y: **Recovering metabolic networks**

using a novel hyperlink prediction method. 2016. arXiv:1610.06941.

27. Oyetunde T, Zhang M, Chen Y, Tang Y, Lo C: **BoostGAPFILL: improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods**. *Bioinformatics* 2016, **33**:btw684.

28. Biggs MB, Papin JA: **Metabolic network-guided binning of metagenomic sequence fragments**. *Bioinformatics* 2016, **32**:867–874.

29. Plata G, Fuhrer T, Hsiao T-L, Sauer U, Vitkup D: **Global probabilistic annotation of metabolic networks enables enzyme discovery**. *Nat. Chem. Biol.* 2012, **8**:848–854.

30. Chitale M, Khan IK, Kihara D: **Missing gene identification using functional coherence scores**. *Sci. Rep.* 2016, **6**:31725.

31. Vitkin E, Shlomi T: **MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks**. *Genome Biol.* 2012, **13**:R111.

32. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND: **Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models**. *PLoS Comput. Biol.* 2014, **10**:e1003882.

33. King B, Farrah T, Richards M, Mundy M, Simeonidis E, Price ND: **ProbAnnoWeb and ProbAnnoPy: probabilistic annotation and gap-filling of metabolic reconstructions**. *bioRxiv,* 2017. 151258.

34. Krumholz EW, Libourel IGL: **Sequence-based network completion reveals the**

integrality of missing reactions in metabolic networks. *J. Biol. Chem.* 2015, **290**:19197–19207.

35. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely S a, Palsson BØ, Agarwalla S: **Experimental and computational assessment of conditionally essential genes in *Escherichia coli*.** *J. Bacteriol.* 2006, **188**:8259–71.

36. Szappanos B, Kovács K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, Gelius-Dietrich G, Lercher MJ, Jelasity M, Myers CL, et al.: **An integrated approach to characterize genetic interaction networks in yeast metabolism**. *Nat. Genet.* 2011, **43**:656–662.

37. Shea A, Wolcott M, Daefler S, Rozak DA: **Biolog phenotype microarrays**. In *Microbial Systems Biology: Methods and Protocols*. Edited by Navid A. Humana Press; 2012:331–373.

38. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, Blow MJ, Bristow J, Butland G, Arkin AP, et al.: **Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons**. *mBio* 2015, **6**:e00306-15.

39. Gallagher RR, Li Z, Lewis AO, Isaacs FJ: **Rapid editing and evolution of bacterial genomes using libraries of synthetic DNA**. *Nat. Protoc.* 2014, **9**:2301–2316.

40. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM: **Programming cells by multiplex genome engineering and accelerated evolution**. *Nature* 2009, **460**:894–898.

41. Nyerges Á, Csörgő B, Nagy I, Bálint B, Bihari P, Lázár V, Apjok G, Umenhoffer K, Bogos B, Pósfai G, et al.: **A highly precise and portable genome engineering method allows comparison of mutational effects across bacterial species**. *Proc Natl Acad Sci USA* 2016, **113**:2502–2507.

42. Ronda C, Pedersen LE, Sommer MOA, Nielsen AT: **CRMAGE: CRISPR optimized MAGE recombineering**. *Sci. Rep.* 2016, **6**:19452.

43. Côté J, French S, Gehrke SS, MacNair CR, Mangat CS, Bharat A, Brown ED: **The genome-wide interaction network of nutrient stress genes in Escherichia coli**. *mBio* 2016, **7**:e01714-16.

44. Giaever G, Nislow C: **The yeast deletion collection: a decade of functional genomics**. *Genetics* 2014, **197**:451–465.

45. Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, Furubayashi A, Kinjyo S, Dose H, Hasegawa M, et al.: **Update on the Keio collection of *Escherichia coli* single-gene deletion mutants.** *Mol. Syst. Biol.* 2009, **5**:335.

46. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H: **Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection.** *Mol. Syst. Biol.* 2006, **2**:2006.0008.

47. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, et al.: **Essential *Bacillus subtilis* genes**. *Proc Natl Acad Sci USA* 2003, **100**:4678–4683.

48. Deutschbauer A, Price MN, Wetmore KM, Shao W, Baumohl JK, Xu Z, Nguyen M, Tamse R, Davis RW, Arkin AP: **Evidence-based annotation of gene function in *Shewanella oneidensis* MR-1 using genome-wide fitness profiling across 121 conditions**. *PLoS Genet.* 2011, **7**:e1002385.

49. Skerker JM, Leon D, Price MN, Mar JS, Tarjan DR, Wetmore KM, Deutschbauer AM, Baumohl JK, Bauer S, Ibanez AB, et al.: **Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates**. *Mol. Syst. Biol.* 2014, **9**:674–674.

50. Dunn AK: ***Vibrio fischeri* metabolism: symbiosis and beyond.** *Adv. Microb. Physiol.* 2012, **61**:37–68.

51. Hastings J, Greenberg EP: **Quorum sensing: The explanation of a curious phenomenon reveals a common characteristic of bacteria**. *J. Bacteriol.* 1999, **181**:2667–2668.

52. Boettcher KJ, Ruby EG: **Depressed light emission by symbiotic *Vibrio fischeri* of the sepiolid squid *Euprymna scolopes***. *J. Bacteriol.* 1990, **172**:3701–3706.

53. Thompson LR, Nikolakakis K, Pan S, Reed J, Knight R, Ruby EG: **Transcriptional characterization of *Vibrio fischeri* during colonization of juvenile *Euprymna scolopes***. *Environ. Microbiol.* 2017, **19**:1845–1856.

54. Brooks JF, Gyllborg MC, Cronin DC, Quillin SJ, Mallama CA, Foxall R, Whistler C, Goodman AL, Mandel MJ: **Global discovery of colonization determinants in the squid symbiont *Vibrio fischeri*.** *Proc Natl Acad Sci USA* 2014, **111**:17284–9.

55. Bongrand C, Koch EJ, Moriano-Gutierrez S, Cordero OX, McFall-Ngai M, Polz MF, Ruby EG: **A genomic comparison of 13 symbiotic *Vibrio fischeri* isolates from the perspective of their host source and colonization behavior.** *ISME J.* 2016, doi:10.1038/ismej.2016.69.

56. Nikolakakis K, Lehnert E, McFall-Ngai MJ, Ruby EG: **Use of hybridization chain reaction-fluorescent *in situ* hybridization to track gene expression by both partners during initiation of symbiosis**. *Appl. Environ. Microbiol.* 2015, **81**:4728–4735.

57. Kim HU, Kim SY, Jeong H, Kim TY, Kim JJ, Choy HE, Yi KY, Rhee JH, Lee SY: **Integrative genome-scale metabolic analysis of *Vibrio vulnificus* for drug targeting and discovery.** *Mol. Syst. Biol.* 2011, **7**:460.

58. Pan S, Nikolakakis K, Adamczyk PA, Pan M, Ruby EG, Reed JL: **Model-enabled gene search (MEGS) allows fast and direct discovery of enzymatic and transport gene functions in the marine bacterium *Vibrio fischeri***. *J. Biol. Chem.* 2017, **292**:10250–10261.

59. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO: **Systems approach to refining genome annotation**. *Proc Natl Acad Sci USA* 2006, **103**:17480–17484.

60. NCBI Resource Coordinators: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 2017, **45**:D12–D17.

61. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, et al.: **The MetaCyc database of**

metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2016, **44**:D471-80.

62. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Res.* 2016, **44**:D457–D462.

63. Textor S, Wendisch VF, De Graaf AA, Müller U, Linder MI, Linder D, Buckel W: **Propionate oxidation in *Escherichia coli*: evidence for operation of a methylcitrate cycle in bacteria.** *Arch. Microbiol.* 1997, **168**:428–36.

64. Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, Kazanov MD, Riehl W, Arkin AP, Dubchak I, et al.: **RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria.** *BMC Genomics* 2013, **14**:745.

65. Nikolakakis K, Monfils K, Moriano-Gutierrez S, Brennan CA, Ruby EG: **Characterization of the *Vibrio fischeri* fatty acid chemoreceptors, VfcB and VfcB2**. *Appl. Environ. Microbiol.* 2016, **82**:696–704.

66. Nyholm S V., Stabb E V., Ruby EG, McFall-Ngai MJ: **Establishment of an animal-bacterial association: recruiting symbiotic vibrios from the environment**. *Proc. Natl Acad Sci USA* 2000, **97**:10231–10235.

67. Mandel MJ, Schaefer AL, Brennan CA, Heath-Heckman EAC, DeLoney-Marino CR, McFall-Ngai MJ, Ruby EG: **Squid-derived chitin oligosaccharides are a chemotactic signal during colonization by *Vibrio fischeri***. *Appl. Environ. Microbiol.* 2012, **78**:4620–4626.

68. Schwartzman JA, Koch E, Heath-Heckman EAC, Zhou L, Kremer N, McFall-Ngai MJ, Ruby EG: **The chemistry of negotiation: rhythmic, glycan-driven acidification in a symbiotic conversation**. *Proc Natl Acad Sci USA* 2015, **112**:566–571.

69. Tran QH, Bongaerts J, Vlad D, Unden G: **Requirement for the proton-pumping NADH dehydrogenase I of *Escherichia coli* in respiration of NADH to fumarate and its bioenergetic implications.** *Eur. J. Biochem.* 1997, **244**:155–60.

70. Lopez O, Morera C, Miranda-Rios J, Girard L, Romero D, Soberón M: **Regulation of gene expression in response to oxygen in *Rhizobium etli*: role of FnrN in fixNOQP expression and in symbiotic nitrogen fixation.** *J. Bacteriol.* 2001, **183**:6999–7006.

71. Peters A, Kulajta C, Pawlik G, Daldal F, Koch H-G: **Stability of the cbb3-type cytochrome oxidase requires specific CcoQ-CcoP interactions**. *J. Bacteriol.* 2008, **190**:5576–86.

72. Pan M, Schwartzman JA, Dunn AK, Lu Z, Ruby EG: **A single host-derived glycan impacts key regulatory nodes of symbiont metabolism in a coevolved mutualism**. *mBio* 2015, **6**:e00811-15.

73. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al.: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res.* 2016, **44**:D733–D745.

74. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM,

Disz T, Gabbard JL, et al.: **Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center.** *Nucleic Acids Res.* 2017, **45**:D535–D542.

75. **UniProtKB: the universal protein knowledgebase**. *Nucleic Acids Res*. 2017, 45:D158-169.

76. Post DMB, Yu L, Krasity BC, Choudhury B, Mandel MJ, Brennan C a, Ruby EG, McFall-Ngai MJ, Gibson BW, Apicella M a: **O-antigen and core carbohydrate of *Vibrio fischeri* lipopolysaccharide: composition and analysis of their role in Euprymna scolopes light organ colonization.** *J. Biol. Chem.* 2012, **287**:8515–30.

77. Nyholm S V.: **Peptidoglycan Monomer Release and *Vibrio fischeri*.** *J. Bacteriol.* 2009, **191**:1997–1999.

78. Buist G, Steen A, Kok J, Kuipers OP: **LysM, a widely distributed protein motif for binding to (peptido)glycans.** *Mol. Microbiol.* 2008, **68**:838–847.

79. Brooks TM, Unterweger D, Bachmann V, Kostiuk B, Pukatzki S: **Lytic activity of the *Vibrio cholerae* type VI secretion toxin VgrG-3 Is inhibited by the antitoxin TsaB.** *J. Biol. Chem.* 2013, **288**:7618–7625.

80. Elderkin S, Jones S, Schumacher J, Studholme D, Buck M: **Mechanism of Action of the *Escherichia coli* phage shock protein PspA in repression of the AAA family transcription factor PspF**. *J. Mol. Biol.* 2002, **320**:23–37.

81. Wolfe AJ, Millikan DS, Campbell JM, Visick KL: ***Vibrio fischeri* σ54 controls**

motility, biofilm formation, luminescence, and colonization. *Appl. Environ. Microbiol.* 2004, **70**:2520–2524.

82. Lyell NL, Dunn AK, Bose JL, Stabb E V.: **Bright mutants of *Vibrio fischeri* ES114 reveal conditions and regulators that control bioluminescence and expression of the *lux* operon**. *J. Bacteriol.* 2010, **192**:5103–5114.

83. Miyashiro T, Ruby EG: **Shedding light on bioluminescence regulation in *Vibrio fischeri*.** *Mol. Microbiol.* 2012, **84**:795–806.

84. Karl MD, Nealson HK: **Regulation of cellular metabolism during synthesis and expression of the luminous system in Beneckea and Photobacterium**. *J. Gen. Microbiol.* 1980.

85. Bose JL, Rosenberg CS, Stabb E V.: **Effects of *luxCDABEG* induction in *Vibrio fischeri*: enhancement of symbiotic colonization and conditional attenuation of growth in culture**. *Arch. Microbiol.* 2008, **190**:169–183.

86. Hastings JW: **Bioluminescence: from chemical bonds to photons.** *Ciba Found. Symp.* 1975.

87. Nealson K, Hastings J: **Bacterial bioluminescence: its control and ecological significance.** *Microbiol. Rev.* 1979, **43**:496–518.

88. Kremer N, Philipp EER, Carpentier MC, Brennan CA, Kraemer L, Altura MA, Augustin R, Häsler R, Heath-Heckman EAC, Peyer SM, et al.: **Initial symbiont contact orchestrates host-organ-wide transcriptional changes that prime tissue colonization**. *Cell Host Microbe* 2013, **14**:183–194.

89. Schwartzman J a, Koch E, Heath-Heckman E a C, Zhou L, Kremer N, McFall-Ngai MJ, Ruby EG: **The chemistry of negotiation: rhythmic, glycan-driven acidification in a symbiotic conversation.** *Proc Natl Acad Sci USA.* 2014, doi:10.1073/pnas.1418580112.

90. Deloney-marino CR, Wolfe AJ, Visick KL: **Chemoattraction of *Vibrio fischeri* to serine , nucleosides , and *N*-acetylneuraminic acid , a component of squid light-organ mucus**. *Appl. Environ. Microbiol.* 2003, **69**:7527–7530.

91. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al.: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102**:15545–50.

92. Stancik LM, Stancik DM, Schmidt B, Barnhart DM, Yoncheva YN, Slonczewski JL: **pH-dependent expression of periplasmic proteins and amino acid catabolism in *Escherichia coli***. *J. Bacteriol.* 2002, **184**:4246–4258.

93. Ye Y, Godzik A: **FATCAT: a web server for flexible structure comparison and structure similarity searching**. *Nucleic Acids Res* 2004, **32**:W582-5.

94. Guda C, Lu S, Scheeff ED, Bourne PE, Shindyalov IN: **CE-MC: A multiple protein structure alignment server**. *Nucleic Acids Res.* 2004, **32**.

95. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci USA* 1999, **96**:2896–2901.

96. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86–90.

97. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285–8.

98. Deutschbauer A, Price MN, Wetmore KM, Tarjan DR, Xu Z, Shao W, Leon D, Arkin AP, Skerker JM: **Towards an informative mutant phenotype for every bacterial gene**. *J. Bacteriol.* 2014, **196**:3643–3655.

99. Green ML, Karp PD: **A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.** *BMC Bioinformatics* 2004, **5**:76.

100. Galperin MY, Koonin E V.: **From complete genome sequence to "complete" understanding?** *Trends Biotechnol.* 2010, **28**:398–406.

101. Golyshev MA, Korotkov E V.: **Developing of the computer method for annotation of bacterial genes**. *Adv. Bioinformatics* 2015, **2015**:1–9.

102. UniProt Consortium: **UniProt: a hub for protein information.** *Nucleic Acids Res.* 2015, **43**:D204-12.

103. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation error in public databases: misannotation of molecular function in enzyme superfamilies**. *PLoS Comput. Biol.* 2009, **5**.

104. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, et al.: **A large-scale evaluation of computational**

protein function prediction. *Nat. Methods* 2013, **10**:221–227.

105. Satish Kumar V, Dasika MS, Maranas CD: **Optimization based automated curation of metabolic reconstructions**. *BMC Bioinformatics* 2007, **8**:212.

106. Kumar VS, Maranas CD: **GrowMatch: an automated method for reconciling *in silicolin vivo* growth predictions**. *PLoS Comput. Biol.* 2009, **5**.

107. Green ML, Karp PD: **Using genome-context data to identify specific types of functional associations in pathway/genome databases**. *Bioinformatics* 2007, **23**:i205–i211.

108. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL: **High-throughput generation, optimization and analysis of genome-scale metabolic models.** *Nat. Biotechnol.* 2010, **28**:977–982.

109. Chen L, Vitkup D: **Predicting genes for orphan metabolic activities using phylogenetic profiles.** *Genome Biol.* 2006, **7**:R17.

110. Varaljay VA, Satagopan S, North JA, Witte B, Dourado MN, Anantharaman K, Arbing MA, McCann SH, Oremland RS, Banfield JF, et al.: **Functional metagenomic selection of ribulose 1, 5-bisphosphate carboxylase/oxygenase from uncultivated bacteria**. *Environ. Microbiol.* 2016, **18**:1187–1199.

111. Simon C, Herath J, Rockstroh S, Daniel R: **Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice**. *Appl. Environ. Microbiol.* 2009, **75**:2964–2968.

112. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, Dantas G:

Bacterial phylogeny structures soil resistomes across habitats. *Nature* 2014, **509**:612–6.

113. Lee KH, Ruby EG: **Effect of the squid host on the abundance and distribution of symbiotic *Vibrio fischeri* in nature**. *Appl. Environ. Microbiol.* 1994, **60**:1565–1571.

114. Dunn AK: ***Vibrio fischeri metabolism: symbiosis and beyond***. *Adv. Microb. Physiol.* 2012, **61**: 37–68.

115. Orth JD, Conrad TM, Na J, Lerman J a, Nam H, Feist AM, Palsson BØ: **A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011.** *Mol. Syst. Biol.* 2011, **7**:535.

116. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat. Biotechnol.* 2010, **28**:245–248.

117. Derst C, Henseling J, Röhm KH: **Engineering the substrate specificity of *Escherichia coli* asparaginase.** *Protein Sci.* 2000, **9**:2009–17.

118. Masters PS, Hong JS: **Genetics of the glutamine transport system in Escherichia coli**. *J. Bacteriol.* 1981, **147**:805–819.

119. Nohno T, Saito T, Hong J shiang: **Cloning and complete nucleotide sequence of the Escherichia coli glutamine permease operon (glnHPQ)**. *MGG Mol. Gen. Genet.* 1986, **205**:260–269.

120. Neidhardt FC et al. (Eds): *Escherichia coli* and Salmonella : Cellular and Molecular Biology. American Society for Microbiology; 1996.

121. Quay SC, Dick TE, Oxender DL: **Role of transport systems in amino acid metabolism: leucine toxicity and the branched-chain amino acid transport systems.** *J. Bacteriol.* 1977, **129**:1257–1265.

122. Capitani G, De Biase D, Aurizi C, Gut H, Bossa F, Grütter MG: **Crystal structure and functional analysis of *Escherichia coli* glutamate decarboxylase.** *EMBO J.* 2003, **22**:4027–4037.

123. Shearer AG, Altman T, Rhee CD: **Finding sequences for over 270 orphan enzymes**. *PLoS One* 2014, **9**.

124. Swainston N, Smallbone K, Mendes P, Kell D, Paton N: **The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks.** *J. Integr. Bioinform.* 2011, **8**:186.

125. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J: **The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*.** *PLoS Comput. Biol.* 2013, **9**:e1002980.

126. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, et al.: **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.** *Brief. Bioinform.* 2010, **11**:40–79.

127. Ong WK, Vu TT, Lovendahl KN, Llull JM, Serres MH, Romine MF, Reed JL: **Comparisons of Shewanella strains based on genome annotations, modeling, and experiments.** *BMC Syst. Biol.* 2014, **8**:31.

128. Fondi M, Maida I, Perrin E, Mellera A, Mocali S, Parrilli E, Tutino ML, Li?? P, Fani R: **Genome-scale metabolic reconstruction and constraint-based modelling of the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125**. *Environ. Microbiol.* 2015, **17**:751–766.

129. Johnston JW, Zaleski A, Allen S, Mootz JM, Armbruster D, Gibson BW, Apicella MA, Munson RS: **Regulation of sialic acid transport and catabolism in *Haemophilus influenzae***. *Mol. Microbiol.* 2007, **66**:26–39.

130. Gaida SM, Sandoval NR, Nicolaou SA, Chen Y, Venkataramanan KP, Papoutsakis ET: **Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries.** *Nat. Commun.* 2015, **6**:7045.

131. Herring CD, Blattner FR: **Conditional lethal amber mutations in essential *Escherichia coli* genes**. *J. Bacteriol.* 2004, **186**:2673–2681.

132. Yang S, Fei Q, Zhang Y, Contreras LM, Utturkar SM, Brown SD, Himmel ME, Zhang M: ***Zymomonas mobilis* as a model system for production of biofuels and biochemicals**. *Microb. Biotechnol.* 2016, **9**:699–717.

133. He MX, Wu B, Qin H, Ruan ZY, Tan FR, Wang JL, Shui ZX, Dai LC, Zhu QL, Pan K, et al.: ***Zymomonas mobilis*: a novel platform for future biorefineries**. *Biotechnol Biofuels* 2014, **7**:101.

134. Lee KY, Park JM, Kim TY, Yun H, Lee SY: **The genome-scale metabolic network analysis of *Zymomonas mobilis* ZM4 explains physiological features and suggests ethanol and succinic acid production strategies**. *Microb Cell Fact*

2010, **9**:94.

135. Widiastuti H, Kim JY, Selvarasu S, Karimi IA, Kim H, Seo JS, Lee DY: **Genome-scale modeling and *in silico* analysis of ethanologenic bacteria *Zymomonas mobilis***. *Biotechnol. Bioeng.* 2011, **108**:655–665.

136. Machado D, Herrgård M: **Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism.** *PLoS Comput. Biol.* 2014, **10**:e1003580.

137. Graf J, Dunlap P V., Ruby EG: **Effect of transposon-induced motility mutations on colonization of the host light organ by *Vibrio fischeri***. *J. Bacteriol.* 1994, **176**:6986–6991.

138. Long CP, Antoniewicz MR: **Quantifying biomass composition by gas chromatography/mass spectrometry.** *Anal. Chem.* 2014, **86**:9423–7.

139. Fritzemeier CJ, Hartleb D, Szappanos B, Papp B, Lercher MJ: **Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal**. *PLOS Comput. Biol.* 2017, **13**:e1005494.

140. Stabb E V., Butler MS, Adin DM: **Correlation between osmolarity and luminescence of symbiotic *Vibrio fischeri* strain ES114**. *J. Bacteriol.* 2004, **186**:2906–2908.

141. Robinson MD, McCarthy DJ, Smyth GK: **EdgeR: a bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139–40.

142. Datsenko KAA, Wanner BLL: **One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products.** *Proc Natl Acad Sci U S A* 2000, **97**:6640–5.

143. Stabb E V., Ruby EG: **RP4-based plasmids for conjugation between *Escherichia coli* and members of the Vibrionaceae**. *Methods Enzymol.* 2002, **358**:413–426.

144. Le Roux F, Binesse J, Saulnier D, Mazel D: **Construction of a *Vibrio splendidus* mutant lacking the metalloprotease gene vsm by use of a novel counterselectable suicide vector**. *Appl. Environ. Microbiol.* 2007, **73**:777–784.

145. Shibata S, Visick KL: **Sensor kinase RscS induces the production of antigenically distinct outer membrane vesicles that depend on the symbiosis polysaccharide locus in *Vibrio fischeri***. *J. Bacteriol.* 2012, **194**:185–194.

146. Dunn AK, Millikan DS, Adin DM, Bose JL, Stabb E V.: **New rfp- and pES213-derived tools for analyzing symbiotic *Vibrio fischeri* reveal patterns of infection and lux expression *in situ***. *Appl. Environ. Microbiol.* 2006, **72**:802–810.

147. Neidhardt FC, Bloch PL, Smith DF: **Culture medium for Enterobacteria**. *J. Bacteriol.* 1974, **119**:736–747.

148. Hayes CS, Bose B, Sauer RT: **Proline residues at the C terminus of nascent chains induce SsrA tagging during translation termination**. *J. Biol. Chem.* 2002, **277**:33825–33832.

149. Miranda-Santos I, Gramacho S, Pineiro M, Martinez-Gomez K, Fritz M, Hollemeyer

K, Salvador A, Heinzle E: **Mass isotopomer analysis of nucleosides isolated from RNA and DNA using GC-MS.** *Anal. Chem.* 2015, **87**:617–623.

150. Millard P, Letisse F, Sokol S, Portais JC: **IsoCor: correcting MS data in isotope labeling experiments**. *Bioinformatics* 2012, **28**:1294–1296.

151. Liu YC, Hsu DH, Huang CL, Liu YL, Liu GY, Hung HC: **Determinants of the differential antizyme-binding affinity of ornithine decarboxylase**. *PLoS One* 2011, **6**.

152. Hsieh JY, Yang JY, Lin CL, Liu GY, Hung HC: **Minimal Antizyme peptide fully functioning in the binding and inhibition of ornithine decarboxylase and Antizyme inhibitor**. *PLoS One* 2011, **6**.

153. Su KL, Liao YF, Hung HC, Liu GY: **Critical factors determining dimerization of human antizyme inhibitor**. *J. Biol. Chem.* 2009, **284**:26768–26777.

154. Jackson LK, Goldsmith EJ, Phillips MA: **X-ray structure determination of Trypanosoma brucei ornithine decarboxylase bound to D-ornithine and to G418 insights into substrate binding and odc conformational flexibility**. *J. Biol. Chem.* 2003, **278**:22037–22043.

155. Liao C, Wang Y, Tan X, Sun L, Liu S: **Discovery of novel inhibitors of human S-adenosylmethionine decarboxylase based on in silico high-throughput screening and a non-radioactive enzymatic assay.** *Sci. Rep.* 2015, **5**:10754.

156. Schefe JH, Lehmann KE, Buschmann IR, Unger T, Funke-Kaiser H: **Quantitative real-time RT-PCR data analysis: current concepts and the novel “gene**

expression's C (T) difference" formula. *J Mol Med* 2006, **84**:901–910.

157. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H: **Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection**. *Mol. Syst. Biol.* 2006, **2**:2006.0008.

158. Sternberg N, Hoess R: **The molecular genetics of bacteriophage P1**. *Annu Rev Genet* 1983, **17**:123–54.